

Groups of Repairmen and Repair-based Load Balancing in Supermarket Models with Repairable Servers

Na Li

Department of Industrial Engineering and Management
Shanghai Jiaotong University, Shanghai 200240, China

Quan-Lin Li

School of Economics and Management Sciences
Yanshan University, Qinhuangdao 066004, China

Zhe George Zhang

Department of Decision Sciences, Western Washington University,
Beedie School of Business, Simon Fraser University

March 28, 2017

Abstract

Supermarket models are a class of interesting parallel queueing networks with dynamic randomized load balancing and real-time resource management. When the parallel servers are subject to breakdowns and repairs, analysis of such a supermarket model becomes more difficult and challenging. In this paper, we apply the mean-field theory to studying four interrelated supermarket models with repairable servers, and numerically indicate impact of the different repairman groups on performance of the systems. First, we set up the systems of mean-field equations for the supermarket models with repairable servers. Then we prove the asymptotic independence of the supermarket models through the operator semigroup and the mean-field limit. Furthermore, we show that the fixed points of the supermarket models satisfy the systems of nonlinear equations. Finally, we use the fixed points to give numerical computation for performer analysis, and provide valuable observations on model improvement. Therefore, this paper provides a new and effective method in the study of complex supermarket models.

Keywords: Supermarket model; mean-field theory; fixed point; performance analysis; repairable server; reliability.

1 Introduction

In the last two decades considerable research attention has been paid to the study of supermarket models. Supermarket models are a class of interesting parallel queueing networks with dynamic and real-time adaptive control, for example, size-based routine selection, and information-based resource scheduling. Such a supermarket model can be applied to, such as, computer networks, manufacturing systems, transportation networks and healthcare systems. Since a simple supermarket model was discussed by Mitzenmacher [43], Vvedenskaya et al. [55] and Turner [53, 54], more studies have been done by, for instance, Vvedenskaya and Suhov [56], Graham [21, 22], Luczak and McDiarmid [37, 38], Bramson et al. [5, 9, 10], Li [28], Li et al. [30, 31] and Li and Lui [32], Gast et al. [19], Gast and Gaujal [20] and Mukhopadhyay et al. [45]. For the fast Jackson networks (or supermarket networks), readers may refer to Martin and Suhov [40], Martin [39] and Suhov and Vvedenskaya [51].

In many stochastic networks, servers subject to breakdowns and repairs always encounter in practical areas, such as, computer systems, communication networks, manufacturing systems, and transportation networks. Because system performance deteriorates quickly due to servers' breakdowns and limited repair capacity, analyzing such a stochastic systems with repairable servers is not only important from theoretical perspective but also necessary from practical engineering. On this research line, important examples include Mitrany and Avi-Ttzhak [42], Neuts and Lucantoni [46], Kulkarni and Choi [26], Li et al. [36], Aissani and Artalejo [2], Núñez-Queija [47], Li et al. [34], Economou and Kantaa [15], Fiems et al. [17], Kamoun [23], and Krishnamoorthy et al. [25] for a survey.

It is interesting but difficult to discuss stochastic systems of parallel queues with unreliable servers, e.g., see an excellent survey by Adan et al. [1]. Up to now, the available results on parallel-queue systems with repairable servers are still very few. Andradottir et al. [3] applied a Markov decision process to compensating for failures with flexible servers. Martonosi [41] studied a dynamic server allocation at parallel queues with unreliable servers. Saghaian et al. [49] analyzed the dynamic control of unreliable flexible servers in a “W” network. Ravid et al. [48] considered the repair systems with exchange-

able items and the longest queue mechanism. Stimulated by practical need of many distributed parallel systems, the study of supermarket models and work stealing models is highly paid attention on computer systems and communication networks. This motivates us in this paper to apply the mean-field theory to analyzing supermarket models with servers subject to breakdowns and repairs, which are a class of important complex reliability networks, and specifically, the different groups of repairmen make analysis of such a reliability network more difficult and challenging.

It is necessary to provide a simple survey for the mean-field theory. The mean-field equations and the asymptotic independence (or propagation of chaos) play an important role in the study of interacting particle systems, e.g., see Liggett [35] and Kipnis and Landim [24]. For the mean-field theory of complex stochastic systems, readers may refer to, for example, interacting Markov processes by Spitzer [50], Dawson [12], Sznitman [52] and Chen [11] and Li [29]; queueing networks by Baccelli et al. [4], Borovkov [7] and Mitzenmacher et al. [44]; work stealing models by Gast and Gaujal [18] and Li and Yang [33]; communication networks by Duffield [13], Benaïm and Le Boudec [6], Duffy [14] and Bordenave et al. [8].

The main contributions of this paper are threefold. The first one is to describe and analyze a class of important complex reliability networks: Supermarket models with repairable servers, which play a key role in performance evaluation of computer systems and of communication networks. Notice that a supermarket model contains multiple repairable servers, thus the different groups of repairmen make analysis of the supermarket model more complicated. In the situation, this paper considers four interrelated supermarket models with repairable servers through observing two different arrival dispatched schemes and two different groups of repairmen. The second contribution is to apply the mean-field theory to studying the four interrelated supermarket models with repairable servers. This paper demonstrates such a mean-field analysis through the following three steps: (a) Providing a probability computation for setting up the systems of mean-field equations, (b) calculating the fixed points through the systems of nonlinear equations, and (c) giving performance analysis of the supermarket models with repairable servers and developing numerical computation for useful observation on model improvement. The third contribution is to provide a better example in order to demonstrate how to develop numerical solution in the study of complex supermarket models. Since the nonlinear structure of the mean-field equations makes a supermarket model almost impossible to find an an-

alytic solution to the system of mean-field equations, it is a key to sufficiently develop numerical computation in performance evaluation of supermarket models. Based on this, numerical examples are used to provide valuable observations on how to improve performance of supermarket models either from system parameter optimization or from various resource deployment (e.g., arrival dispatched schemes, allocated service ability, and groups of repairmen).

Finally, note that this paper discusses a special class of supermarket models with unreliable servers, while their failed states and the groups of repairmen have influence on the arrival joining schemes. To analyze such a supermarket model, the most relevant references to this paper are Li et al. [30, 31] and Li and Lui [32] from two points of view: (1) The environment invariant factors were proposed to setting up systems of mean-field equations for complex supermarket models. As studied in Li et al. [31], this paper also analyzes a double dynamic routine selection scheme both for the arrival dispatched schemes and for the groups of repairmen. It is worthwhile to note that such a multiple dynamic routine selection scheme is a new and interesting topic in the study of supermarket models and of work stealing models.

The remainder of this paper is organized as follows. In Section 2, we first describe four interrelated supermarket models with repairable servers where customer arrivals make use of system information and repair ability is grouped in some different structures. Then we use the fraction vector to describe an infinite-dimensional Markov process for each supermarket model with repairable servers. In Sections 3, we provide two types of probability representations both for the arrival dispatched schemes by means of system information and for the repair ability grouped in different ways. In Sections 4, for each of the four interrelated supermarket models with repairable servers, we set up an infinite-dimensional system of mean-field equations. In Section 5, we discuss the fixed points for the systems of mean-field equations, and show that the fixed points can be determined by the systems of nonlinear equations. In Section 6, we first provide useful performance measures of the supermarket models with repairable servers. Then we use some numerical examples to make valuable observations on model improvement by means of performance numerical comparison. Section 7 concludes with a summary. The proofs of some key results are provided in Appendix A.

2 Supermarket Models with Repairable Servers

In this section, we first describe four interrelated supermarket models with repairable servers, where the arrival dispatched schemes make use of system information and the repair ability is grouped in different ways. Then we use the fraction vector (or empirical measure) to describe an infinite-dimensional Markov process for each supermarket model with repairable servers.

2.1 Model description

The arrival processes

Customers arrive at the system as a Poisson process with arrival rate $N\lambda$ for $\lambda > 0$. Upon arrival, an arriving customer chooses $d_1(\geq 1)$ servers from the N servers independently and randomly. Then the customer will select one server (or queue) to join. Such a server selection is based on two different information observations as follows:

(A.1) Observing only the shortest queue. The arriving customer joins the shortest queue among the d_1 queues. If there is a tie, the customer makes the choice equally likely among the shortest queues of the same length.

(A.2) Observing both the shortest queue and the status (working or repairing) of the d_1 selected servers. The arriving customer joins the shortest queue with the working server as higher priority than the server in repair among the d_1 selected servers.

The service processes

The service times at each server are i.i.d. and are exponential distributed with service rate $\mu > 0$.

The repair processes

Each server has an exponential life time with failure rate $\alpha > 0$. When the server fails, it enters a failure state and undergoes the repair process immediately. The service of a customer interrupted by a server's failure is resumed as soon as the server is repaired. We assume that the repaired server is as good as new and the service time is cumulative. To deploy the repair resource effectively, we consider three types of repair schemes as follows:

(R.1) Each server has one repairman. There are N repairmen corresponding to the N servers, and thus each server has a repairman of itself. The repair times are i.i.d exponential random variables with repair rate β .

(R.2) A super large repairman. There is only one fast repairman whose repair time

is exponentially distributed with repair rate $N\beta$ and $\beta > 0$. This super repairman chooses $d_2(\geq 1)$ servers from the N servers randomly. If all the d_2 servers are working, then the repairman is idle; if at least one of the selected d_2 servers is failed, then the repairman repairs the failed server with the longest queue. If there is a tie, the repairman select one randomly.

(R.3) A large repairman and J small repairmen for $0 \leq J \leq N - 1$. There are a large repairman and J small repairmen, where the repair time of the large repairman is exponentially distributed with the repair rate $(N - J)\beta$, and the repair time of each small repairman is exponentially distributed with repair rate β .

Each of the J small repairmen can repair one failed server at a time, if any; whilst the large repairman chooses d_2 servers from the N servers independently and randomly. If all the selected d_2 servers are working, then the large repairman is idle; if at least one of the selected d_2 servers is failed but not repaired by small repairmen yet, then the large repairman repairs the failed server with the longest queue. If there is a tie, the repairman selects the failed server with the longest queue.

We assume that all the random variables defined above are independent of each other. Figure 1 shows a supermarket model with repairable servers and a large repairman.

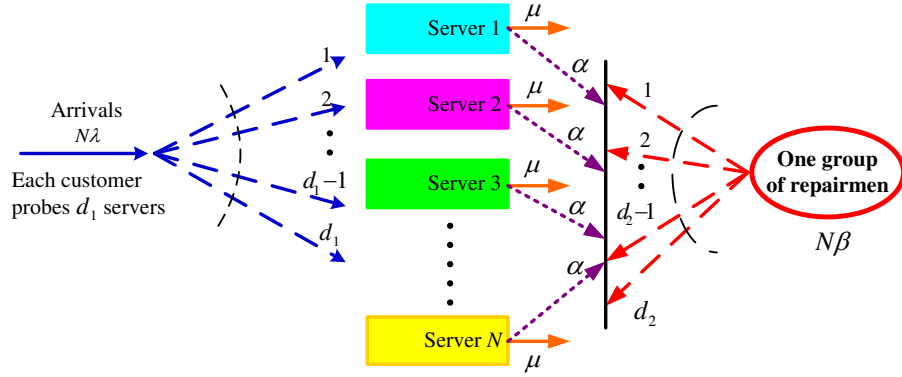


Figure 1: A physical illustration of a supermarket model with repairable servers

Now, we construct four interrelated supermarket models with repairable servers, which are constructed by different combinations of (A.i) and (R.i) for $i = 1, 2$ as follows:

Model I ((A.1) and (R.1)): In this model, an arriving customer only needs to observe the queue lengths of the d_1 selected server and joins the shortest queue. There are N

repairmen corresponding to the N servers, hence each server has one repairman of itself.

Model II ((A.1) and (R.2)): In this model, the queue selection rule is the same as Model I. However, there is only one super repairman who chooses d_2 servers from the N servers independently and uniformly at random. If all the selected d_2 servers are working, then the repairman is idle; otherwise, the repairman repairs the failed server with the longest queue length.

Model III ((A.2) and (R.1)): In this model, an arriving customer observe not only the queue lengths of the d_1 selected servers, but also the states (working or repairing) of the d_1 selected servers. The customer then joins the shortest queue with working servers having higher priority than failed servers. There are N repairmen corresponding to the N servers, hence each server has one repairman of itself.

Model IV ((A.2) and (R.2)): In this model, the customer's queue selection rule is the same as Model III. However, there is only a super repairman, which chooses d_2 servers from the N servers independently and uniformly at random. If all the selected d_2 servers are working, then the repairman is idle; otherwise, the repairman repairs the failed server with the longest queue.

Remark 1 *Actually, (R.3) is a more general scheme of repair resource allocation, and its analysis can be completed through by modifying the mean-field equations in (R.1) and (R.2). Here, we do not consider (R.3), and (R.3) will be investigated in another paper.*

Next, we shall provide a complete mathematical analysis for the four interrelated supermarket models, and present some numerical examples to show how the system information ((A. i) and repair resource allocation (R. i) for $i = 1, 2$) affect performance of the supermarket models with repairable servers. Some insightful observations are made for designing and controlling the arrival, service and repair processes to improve the supermarket models.

2.2 An infinite-dimensional Markov process

Now, we use the empirical measure to provide an infinite-dimensional Markov process for studying each of the four interrelated supermarket models with repairable servers.

For $k \geq 0$, we denote by $n_k^{(W)}(t)$ the numbers of working (or idle) servers with at least $k \geq 0$ customers at time $t \geq 0$, and $n_l^{(R)}(t)$ the numbers of failed servers with at least $l \geq 1$

customers at time $t \geq 0$. Clearly, $n_0^{(W)}(t) + n_1^{(R)}(t) = N$ and $0 \leq n_k^{(W)}(t), n_l^{(R)}(t) \leq N$ for $k \geq 0$ and $l \geq 1$.

We write that for $k \geq 0$,

$$U_{W,k}^{(N)}(t) = \frac{n_k^{(W)}(t)}{N}$$

and for $l \geq 1$

$$V_{R,l}^{(N)}(t) = \frac{n_l^{(R)}(t)}{N},$$

which are the fractions of working (or idle) servers with at least k customers and of failed servers with at least l customers at time $t \geq 0$, respectively. Let

$$\mathbf{U}_W^{(N)}(t) = \left(U_{W,0}^{(N)}(t), U_{W,1}^{(N)}(t), U_{W,2}^{(N)}(t), \dots \right),$$

$$\mathbf{V}_R^{(N)}(t) = \left(V_{R,1}^{(N)}(t), V_{R,2}^{(N)}(t), V_{R,3}^{(N)}(t), \dots \right),$$

and

$$\mathbf{U}^{(N)}(t) = \left(\mathbf{U}_W^{(N)}(t), \mathbf{V}_R^{(N)}(t) \right).$$

Clearly, the state of the supermarket model of N identical repairable servers is described as a stochastic process $\{\mathbf{U}^{(N)}(t) : t \geq 0\}$. Since the arrival process is Poisson, and the distributions of the service, life and repair times are all exponential, $\{\mathbf{U}^{(N)}(t) : t \geq 0\}$ is an infinite-dimensional Markov process whose state space is given by

$$\begin{aligned} \mathbf{E}_N = \left\{ \left(\mathbf{u}^{(N)}, \mathbf{v}^{(N)} \right) : 1 \geq u_0^{(N)} \geq u_1^{(N)} \geq u_2^{(N)} \geq u_3^{(N)} \geq \dots \geq 0, \right. \\ \left. 1 \geq v_1^{(N)} \geq v_2^{(N)} \geq v_3^{(N)} \geq v_4^{(N)} \geq \dots \geq 0, \right. \\ \left. Nu_k^{(N)} \text{ and } Nv_l^{(N)} \text{ are nonnegative integers for } k \geq 0 \text{ and } l \geq 1 \right\}. \end{aligned}$$

For a fixed pair array (t, N) with $t \geq 0$ and $N = 1, 2, 3, \dots$, it is easy to see from the stochastic order that $U_{W,k}^{(N)}(t) \geq U_{W,k+1}^{(N)}(t)$ for $k \geq 0$ and $V_{R,l}^{(N)}(t) \geq V_{R,l+1}^{(N)}(t)$ for $l \geq 1$. This gives

$$1 \geq U_{W,0}^{(N)}(t) \geq U_{W,1}^{(N)}(t) \geq U_{W,2}^{(N)}(t) \geq U_{W,3}^{(N)}(t) \geq \dots \geq 0 \quad (1)$$

and

$$1 \geq V_{R,1}^{(N)}(t) \geq V_{R,2}^{(N)}(t) \geq V_{R,3}^{(N)}(t) \geq V_{R,4}^{(N)}(t) \geq \dots \geq 0. \quad (2)$$

To study the infinite-dimensional Markov process $\{\mathbf{U}^{(N)}(t) : t \geq 0\}$, we write the expected fractions as follows

$$u_{W,k}^{(N)}(t) = E \left[U_{W,k}^{(N)}(t) \right]$$

and

$$u_{R,l}^{(N)}(t) = E \left[V_{R,l}^{(N)}(t) \right].$$

It is easy to see from (1) and (2) that

$$1 \geq u_{W,0}^{(N)}(t) \geq u_{W,1}^{(N)}(t) \geq u_{W,2}^{(N)}(t) \geq u_{W,3}^{(N)}(t) \geq \dots \geq 0 \quad (3)$$

and

$$1 \geq u_{R,1}^{(N)}(t) \geq u_{R,2}^{(N)}(t) \geq u_{R,3}^{(N)}(t) \geq u_{R,4}^{(N)}(t) \geq \dots \geq 0. \quad (4)$$

Let

$$\begin{aligned} \mathbf{u}_W^{(N)}(t) &= \left(u_{W,0}^{(N)}(t), u_{W,1}^{(N)}(t), u_{W,2}^{(N)}(t), u_{W,3}^{(N)}(t), \dots \right), \\ \mathbf{V}_R^{(N)}(t) &= \left(u_{R,1}^{(N)}(t), u_{R,2}^{(N)}(t), u_{R,3}^{(N)}(t), u_{R,4}^{(N)}(t), \dots \right) \end{aligned}$$

and

$$\mathbf{u}^{(N)}(t) = \left(\mathbf{u}_W^{(N)}(t), \mathbf{V}_R^{(N)}(t) \right).$$

3 Two Types of Probability Representations

In this section, we provide two types of probability representations for customer arrivals by means of system information and for repair ability grouped in different ways. For notational simplicity, the two types of probability representations are denoted as the four pair control schemes: $((A.i), (R.i))$ for $i = 1, 2$. The probability representations are useful for establishing the systems of mean-field equations later.

For the supermarket models of N identical repairable servers, to set up the probability representations, we only need to determine the expected change in the number of servers with at least k customers over a small time period $[0, dt)$.

3.1 The arrival processes

This subsection provides the probability representations for the arrival processes, in which the two different cases of (A.1) and (A.2) are discussed. Note that the analysis of (A.1) is similar to that of Li et al. [30]. To make our paper self-contained, we still present some computational details for (A.1) and (R.1). For (A.2) and (R.2), we only provide the main results.

(A.1): Observing the Queue Length Only

To give the probability representations, we need to compute the rate that any arriving customer selects d_1 servers from the N servers independently and uniformly at random, and joins the selected server with the shortest queue. Note that the arriving customer does not have the server status information (working or repair). Thus our computation for such a rate contains two steps as follows:

Step I: Entering one working server

In this step, the rate that any arriving customer joins a working server with the shortest queue length $k - 1$ is given by

$$N\lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) dt, \quad (5)$$

where

$$\begin{aligned} W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) &= \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right]^{m-1} \left[u_{W,1}^{(N)}(t) \right]^{d_1-m} \\ &+ \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right]^{m-1} \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{W,1}^{(N)}(t) \right]^{d_1-m-j} \left[u_{R,1}^{(N)}(t) \right]^j, \end{aligned} \quad (6)$$

and for $k \geq 2$

$$\begin{aligned} W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m-1} \left[u_{W,k}^{(N)}(t) \right]^{d_1-m} \\ &+ \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m-1} \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{W,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{R,k}^{(N)}(t) \right]^j \\ &+ \sum_{m=2}^{d_1} C_{d_1}^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m_1}^{m_1} \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m_1-1} \\ &\times \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-m_1} \sum_{r=0}^{d_1-m} C_{d_1-m}^r \left[u_{W,k}^{(N)}(t) \right]^r \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-r}. \end{aligned} \quad (7)$$

To derive the probabilities $W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$ and $W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ for $k \geq 2$, Figure 2 shows the set decomposition of all possible events, and the probabilities are derived from the following three parts, that is,

$$W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) = \text{Part I} + \text{Part II} + \text{Part III}.$$

Part I: None of the d_1 selected servers is in repair. All d_1 selected servers are working for serving customers. In this case, the probability that any arriving customer joins a

Each of the d_1 selected servers is working for service, and there is at least one working server with the shortest queue length $k-1$. (Part I)	
In the d_1 selected servers, there is at least one working server with the shortest queue length $k-1$, and there exists at least one repaired server while the queue length of each repaired server is more than k customers. (Part II)	In the d_1 selected servers, there are at least one working server with the shortest queue length $k-1$ and at least one repaired server with the shortest queue length $k-1$. (Part III)

Figure 2: Set decomposition of possible events when joining a working server

working server with the shortest queue length $k - 1$ and the queue lengths of the other selected $d_1 - 1$ working servers are not shorter than $k - 1$ is given by

$$\begin{aligned}
& \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^m \left[u_{W,k}^{(N)}(t) \right]^{d_1-m} \\
&= \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m-1} \left[u_{W,k}^{(N)}(t) \right]^{d_1-m}, \quad (8)
\end{aligned}$$

where $C_{d_1}^m = d_1! / [m! (d_1 - m)!]$ is a binomial coefficient, $\left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^m$ is the probability that any arriving customer who can only choose one queue makes m independent selections during the m selected working servers with the queue length $k - 1$ at time t , and $\left[u_{W,k}^{(N)}(t) \right]^{d_1-m}$ is the probability that any arriving customer who can only choose one queue makes $d_1 - m$ independent selections during the $d_1 - m$ selected working servers whose queue lengths are not shorter than k at time t .

Part II: For the d_1 selected servers, there is at least one working server with the shortest queue length $k - 1$, and there exist at least one server in repair while the queue length of each server in repair is more than k customers. In this case, the probability that any arriving customer joins a working server with the shortest queue length $k - 1$; and for the other $d_1 - 1$ selected servers, the queue lengths of the selected working servers are not shorter than $k - 1$, and there exist at least one server in repair while the queue length of

each server in repair is more than k customers, is given by

$$\begin{aligned}
& \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^m \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{W,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{R,k}^{(N)}(t) \right]^j \\
&= \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m-1} \\
&\quad \times \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{W,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{R,k}^{(N)}(t) \right]^j. \tag{9}
\end{aligned}$$

Part III: For the d_1 selected servers, there is at least one working server with the shortest queue length $k-1$ and there is at least one server in repair with the shortest queue length $k-1$. In this case, if there are the m selected servers with the shortest queue length $k-1$ where there are m_1 working servers and $m-m_1$ servers in repair, then the probability that any arriving customer joins a working server is equal to m_1/m . Therefore, the probability that any arriving customer joins a working server with the shortest queue length $k-1$, the queue lengths of the other d_1-1 selected servers are not shorter than $k-1$, and there are at least one working server with $k-1$ customers and at least one server in repair with $k-1$ customers is given by

$$\begin{aligned}
& \sum_{m=2}^{d_1} C_{d_1}^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m_1}^{m_1} \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m_1} \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-m_1} \\
&\quad \times \sum_{r=0}^{d_1-m} C_{d_1-m}^r \left[u_{W,k}^{(N)}(t) \right]^r \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-r} \\
&= \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] \sum_{m=2}^{d_1} C_{d_1}^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m_1}^{m_1} \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m_1-1} \\
&\quad \times \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-m_1} \sum_{r=0}^{d_1-m} C_{d_1-m}^r \left[u_{W,k}^{(N)}(t) \right]^r \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-r}. \tag{10}
\end{aligned}$$

Step two: Entering one server in repair

This step can be dealt with similarly to that in Step one. The rate that any arriving customer joins one server in repair with the shortest queue length $k-1$ and the queue lengths of the other selected d_1-1 servers are not shorter than $k-1$ is given by

$$N\lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) dt, \tag{11}$$

where

$$\begin{aligned}
W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \left[u_{R,k}^{(N)}(t) \right]^{d_1-m} \\
&+ \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{W,k}^{(N)}(t) \right]^j \\
&+ \sum_{m=2}^{d_1} C_{d_1}^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m_1}^{m_1} \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m_1-1} \\
&\times \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right]^{m-m_1} \sum_{r=0}^{d_1-m} C_{d_1-m}^r \left[u_{R,k}^{(N)}(t) \right]^r \left[u_{W,k}^{(N)}(t) \right]^{d_1-m-r}. \tag{12}
\end{aligned}$$

The following theorem simplifies expressions for the probabilities $W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ and $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ for $k \geq 2$, while its proof is similar to that in Theorem 1 of Li et al. [30] and is omitted here. Note that the simplified expressions will be a key in our later study, for example, the system of mean-field equations can be simplified significantly and the fixed point can be computed effectively.

Theorem 1

$$W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) = \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right]^{m-1} \left[u_{W,1}^{(N)}(t) + u_{R,1}^{(N)}(t) \right]^{d_1-m},$$

and for $k \geq 2$

$$\begin{aligned}
W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\
&= \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) + u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \left[u_{W,k}^{(N)}(t) + u_{R,k}^{(N)}(t) \right]^{d_1-m}.
\end{aligned}$$

Using Theorem 1, we set

$$L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) = W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$$

and for $k \geq 2$

$$\begin{aligned}
L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\
&= W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t).
\end{aligned}$$

(A.2): Observing Both the Queue Length and the States of the Chosen Servers

In this case, the arriving customer has a priority for joining one working server with the shortest queue length. Upon arrival, each customer chooses $d_1 \geq 1$ servers from the N servers independently and uniformly at random, and joins the one whose queue length is the shortest among the d_1 servers. If the servers with the shortest queue length contain at least one working server and at least one server in repair, then the arriving customer must randomly join one of the working servers with the shortest queue length. If there is a tie, the working servers with the shortest queue length are chosen randomly.

It is seen that the only difference from (A.1) is that the arriving customer can not join one of the repairing servers with the shortest queue length when there exists at least one working server with the shortest queue length. Based on this, we have

a) The probabilities $W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$ and $W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ for $k \geq 2$ are the same as those in (A.1).

b) Comparing with the probabilities $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ in (A.1), for (A.2) we obtain that for $k \geq 2$

$$\begin{aligned} \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \left[u_{R,k}^{(N)}(t) \right]^{d_1-m} \\ &+ \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{W,k}^{(N)}(t) \right]^j. \end{aligned}$$

Note that Part III of computing $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ in (A.1) is omitted by utilizing the information of (A.2).

3.2 The repair processes

Now we provide the probability representations for the repair processes in two cases: (R.1) and (R.2).

(R.1): Each Server Has One Repairman

In this case, there are N repairmen corresponding to the N servers, hence each server has one repairman. Since the repair time is exponentially distributed with repair rate β , it is seen from Li et al. [34] if the service time of each server is of phase type with irreducible matrix representation (τ, T) , where

$$\tau = (1, 0), \quad T = \begin{pmatrix} -(\mu + \alpha) & \alpha \\ \beta & -\beta \end{pmatrix},$$

then the repairable supermarket model is equivalent to a supermarket model with Poisson inputs and PH service times, as discussed in Li and Lui [28].

(R.2): A Super Repairman

In this case, there is a single super repairman whose repair time is exponentially distributed with repair rate $N\beta$. The repairman chooses d_2 servers from the N servers independently and uniformly at random. If all the selected d_2 servers are in working condition, the repairman is idle; if at least one of the selected d_2 servers is failed, then the repairman attends one failed server with the longest queue. If there is a tie, the repairman select a server randomly.

The rate that the repairman randomly chooses one of the failed servers with the longest queue length k and the queue lengths of the other $d_2 - 1$ selected servers are not longer than k is given by

$$N\beta \sum_{m=1}^{d_2} C_{d_2}^m \left[u_{R,1}^{(N)}(t) \right]^m \left[u_{W,0}^{(N)}(t) - u_{W,2}^{(N)}(t) \right]^{d_2-m} dt \stackrel{\text{def}}{=} N\beta I_1 \left(u_{R,1}^{(N)}, u_{W,0}^{(N)}, u_{W,2}^{(N)}; t \right) dt,$$

and for $k \geq 2$

$$N\beta \sum_{m=1}^{d_2} C_{d_2}^m \left\{ \sum_{m_1=1}^m C_{m_1}^{m_1} \left[u_{R,k}^{(N)}(t) \right]^{m_1} \left[u_{R,1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-m_1} \right\} \left[u_{W,0}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right]^{d_2-m} dt \\ \stackrel{\text{def}}{=} N\beta I_k \left(u_{R,1}^{(N)}, u_{R,k}^{(N)}, u_{W,0}^{(N)}, u_{W,k+1}^{(N)}; t \right) dt.$$

Using $u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1$, we can further simplify

$$I_1 \left(u_{R,1}^{(N)}, u_{W,0}^{(N)}, u_{W,2}^{(N)}; t \right) = \left\{ \left[1 - u_{W,2}^{(N)}(t) \right]^{d_2} - \left[u_{W,0}^{(N)}(t) - u_{W,2}^{(N)}(t) \right]^{d_2} \right\} \\ = \left[1 - u_{W,2}^{(N)}(t) \right]^{d_2} - \left[1 - u_{R,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right]^{d_2} \\ \stackrel{\text{def}}{=} I_1 \left(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}; t \right) \quad (13)$$

and for $k \geq 2$

$$I_k \left(u_{R,1}^{(N)}, u_{R,k}^{(N)}, u_{W,0}^{(N)}, u_{W,k+1}^{(N)}; t \right) = \left[1 - u_{W,k+1}^{(N)}(t) \right]^{d_2} - \left[1 - u_{R,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right]^{d_2} \\ \stackrel{\text{def}}{=} I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right). \quad (14)$$

4 The Mean-Field Equations

In this section, for each of the four interrelated supermarket models with repairable servers, we set up an infinite-dimensional system of mean-field equations. To this end, we present

a detailed analysis only for the first model, while the other three models can be simply discussed on a similar line.

4.1 Model I ((A.1) and (R.1))

For (A.1) and (R.1), the probabilities $W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$, $W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ and $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ for $k \geq 2$ are given in (6), (7) and (12), which are further simplified in Theorem 1.

Now, we consider the service and repair processes. The rate that a customer leaves one server queued by k customers is given by

$$N\mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] dt. \quad (15)$$

The rate that one working server with at least k customers fails is given by

$$N\alpha u_{W,k}^{(N)}(t) dt. \quad (16)$$

The rate that one failed server with at least k customers is repaired is given by

$$N\beta u_{R,k}^{(N)}(t) dt. \quad (17)$$

Based on Equation (5), and Equations (15) to (17), we obtain

$$\begin{aligned} \frac{d}{dt} u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] W_{W,k}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta u_{R,k}^{(N)}(t). \end{aligned} \quad (18)$$

In addition, it follows from (11) that

$$\begin{aligned} \frac{d}{dt} u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] W_{R,k}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta u_{R,k}^{(N)}(t). \end{aligned} \quad (19)$$

Based on the similar analysis to (18) and (19), we can set up an infinite-dimensional system of mean-field equations satisfied by the expected fraction vector $\mathbf{u}^{(N)}(t) = \left(\mathbf{u}_W^{(N)}(t), \mathbf{u}_R^{(N)}(t) \right)$ as follows:

$$\frac{d}{dt} u_{W,0}^{(N)}(t) = -\alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \quad (20)$$

$$\begin{aligned} \frac{d}{dt} u_{W,1}^{(N)}(t) = & \lambda \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right] W_{W,1}(u_{W,0}, u_{W,1}; u_{R,1}; t) \\ & - \mu \left[u_{W,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right] - \alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \end{aligned} \quad (21)$$

for $k \geq 2$

$$\begin{aligned} \frac{d}{dt} u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta u_{R,k}^{(N)}(t) \end{aligned} \quad (22)$$

and

$$\begin{aligned} \frac{d}{dt} u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta u_{R,k}^{(N)}(t), \end{aligned} \quad (23)$$

with the boundary condition

$$u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1, \quad t \geq 0, \quad (24)$$

and the initial conditions

$$\begin{cases} u_{W,k}^{(N)}(0) = g_k, & k \geq 0, \\ u_{R,l}^{(N)}(0) = h_l, & l \geq 1. \end{cases} \quad (25)$$

where

$$1 \geq g_0 \geq g_1 \geq g_2 \geq \cdots \geq 0,$$

$$1 \geq h_1 \geq h_2 \geq h_3 \geq \cdots \geq 0,$$

with

$$g_0 + h_1 = 1.$$

It follows from Theorem 1 that

$$L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) = W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$$

and for $k \geq 2$

$$\begin{aligned} L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) &= W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ &= W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t). \end{aligned} \quad (26)$$

Using $L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$ and $L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ for $k \geq 2$, Equations (20) to (25) can further be simplified as

$$\frac{d}{dt} u_{W,0}^{(N)}(t) = -\alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \quad (27)$$

$$\begin{aligned} \frac{d}{dt} u_{W,1}^{(N)}(t) = & \lambda \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right] L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) \\ & - \mu \left[u_{W,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right] - \alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \end{aligned} \quad (28)$$

for $k \geq 2$

$$\begin{aligned} \frac{d}{dt} u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta u_{R,k}^{(N)}(t) \end{aligned} \quad (29)$$

and

$$\begin{aligned} \frac{d}{dt} u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta u_{R,k}^{(N)}(t), \end{aligned} \quad (30)$$

with the boundary condition

$$u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1, \quad t \geq 0, \quad (31)$$

and the initial conditions

$$\begin{cases} u_{W,k}^{(N)}(0) = g_k, & k \geq 0, \\ u_{R,l}^{(N)}(0) = h_l, & l \geq 1. \end{cases} \quad (32)$$

4.2 Model II ((A.1) and (R.2))

In this model, for (R.2) it follows from (13) and (14) that for $k \geq 1$

$$I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right) = \left[1 - u_{W,k+1}^{(N)}(t) \right]^{d_2} - \left[1 - u_{R,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right]^{d_2}.$$

Hence the dynamic routine selection scheme (R.2) shows that for $k \geq 1$, $\beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right)$ will take the place of $\beta u_{R,k}^{(N)}(t)$ in the systems of mean-field equations (27) to (32). Based on this, we obtain

$$\frac{d}{dt} u_{W,0}^{(N)}(t) = -\alpha u_{W,1}^{(N)}(t) + \beta I_1 \left(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}; t \right), \quad (33)$$

$$\begin{aligned} \frac{d}{dt} u_{W,1}^{(N)}(t) = & \lambda \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right] L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) \\ & - \mu \left[u_{W,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right] - \alpha u_{W,1}^{(N)}(t) + \beta I_1 \left(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}; t \right), \end{aligned} \quad (34)$$

for $k \geq 2$

$$\begin{aligned} \frac{d}{dt} u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right) \end{aligned} \quad (35)$$

and

$$\begin{aligned} \frac{d}{dt} u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right), \end{aligned} \quad (36)$$

with the boundary condition

$$u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1, \quad t \geq 0, \quad (37)$$

and the initial conditions

$$\begin{cases} u_{W,k}^{(N)}(0) = g_k, & k \geq 0, \\ u_{R,l}^{(N)}(0) = h_l, & l \geq 1. \end{cases} \quad (38)$$

4.3 Model III ((A.2) and (R.1))

In this model, the only difference is that an arriving customer cannot join the server in repair with the shortest queue length when there exists at least one working server with the shortest queue length. Thus we obtain

$$\begin{aligned} \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) = & \sum_{m=1}^{d_1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \left[u_{R,k}^{(N)}(t) \right]^{d_1-m} \\ & + \sum_{m=1}^{d_1-1} C_{d_1}^m \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right]^{m-1} \sum_{j=1}^{d_1-m} C_{d_1-m}^j \left[u_{R,k}^{(N)}(t) \right]^{d_1-m-j} \left[u_{W,k}^{(N)}(t) \right]^j \end{aligned} \quad (39)$$

and

$$\begin{aligned} & \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & = \left[u_{R,k-1}^{(N)}(t) + u_{W,k}^{(N)}(t) \right]^{d_1} - \left[u_{W,k}^{(N)}(t) + u_{R,k}^{(N)}(t) \right]^{d_1}. \end{aligned}$$

It is easy to see from (12), (39) and Theorem 1 that

$$\mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \neq L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t). \quad (40)$$

Thus (A.2) indicates that $\mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ needs to replace $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ in the systems of mean-field equations (27) to (32).

On the other hand, except of (40), we still have

$$W_{W,1}^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) = L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$$

and for $k \geq 2$

$$W_{W,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) = L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t).$$

A similar analysis to the systems of mean-field equations (27) to (32), we obtain

$$\frac{d}{dt}u_{W,0}^{(N)}(t) = -\alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \quad (41)$$

$$\begin{aligned} \frac{d}{dt}u_{W,1}^{(N)}(t) = & \lambda \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right] L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) \\ & - \mu \left[u_{W,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right] - \alpha u_{W,1}^{(N)}(t) + \beta u_{R,1}^{(N)}(t), \end{aligned} \quad (42)$$

for $k \geq 2$

$$\begin{aligned} \frac{d}{dt}u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta u_{R,k}^{(N)}(t) \end{aligned} \quad (43)$$

and

$$\begin{aligned} \frac{d}{dt}u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta u_{R,k}^{(N)}(t), \end{aligned} \quad (44)$$

with the boundary condition

$$u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1, \quad t \geq 0, \quad (45)$$

and the initial conditions

$$\begin{cases} u_{W,k}^{(N)}(0) = g_k, & k \geq 0, \\ u_{R,l}^{(N)}(0) = h_l, & l \geq 1. \end{cases} \quad (46)$$

4.4 Model IV ((A.2) and (R.2))

Since (A.2) needs $\mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ replacing $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$, and (R.2) needs $\beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right)$ taking the place of $\beta u_{R,k}^{(N)}(t)$. Thus we obtain

$$\frac{d}{dt}u_{W,0}^{(N)}(t) = -\alpha u_{W,1}^{(N)}(t) + \beta I_1 \left(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}; t \right), \quad (47)$$

$$\begin{aligned} \frac{d}{dt} u_{W,1}^{(N)}(t) = & \lambda \left[u_{W,0}^{(N)}(t) - u_{W,1}^{(N)}(t) \right] L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t) \\ & - \mu \left[u_{W,1}^{(N)}(t) - u_{W,2}^{(N)}(t) \right] - \alpha u_{W,1}^{(N)}(t) + \beta I_1 \left(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}(t) \right), \end{aligned} \quad (48)$$

for $k \geq 2$

$$\begin{aligned} \frac{d}{dt} u_{W,k}^{(N)}(t) = & \lambda \left[u_{W,k-1}^{(N)}(t) - u_{W,k}^{(N)}(t) \right] L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & - \mu \left[u_{W,k}^{(N)}(t) - u_{W,k+1}^{(N)}(t) \right] - \alpha u_{W,k}^{(N)}(t) + \beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right) \end{aligned} \quad (49)$$

and

$$\begin{aligned} \frac{d}{dt} u_{R,k}^{(N)}(t) = & \lambda \left[u_{R,k-1}^{(N)}(t) - u_{R,k}^{(N)}(t) \right] \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t) \\ & + \alpha u_{W,k}^{(N)}(t) - \beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right), \end{aligned} \quad (50)$$

with the boundary condition

$$u_{W,0}^{(N)}(t) + u_{R,1}^{(N)}(t) = 1, \quad t \geq 0, \quad (51)$$

and the initial conditions

$$\begin{cases} u_{W,k}^{(N)}(0) = g_k, & k \geq 0, \\ u_{R,l}^{(N)}(0) = h_l, & l \geq 1. \end{cases} \quad (52)$$

Remark 2 From the four systems of mean-field equations, we find that to set up the systems of mean-field equations, two key rules must be followed as follows:

(1) If (A.1) \rightarrow (A.2), then $\mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$ takes the place of $W_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t)$, and

(2) if (R.1) \rightarrow (R.2), then $\beta I_k \left(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t \right)$ replaces $\beta u_{R,k}^{(N)}(t)$.

5 The Fixed Point

In this section, we discuss the fixed points for the systems of mean-field equations, and show that the fixed points can be determined by the systems of nonlinear equations. Specifically, we indicate that the nonlinear structure makes the analytical solution of the fixed points too complicated and even impossible. Since such a fixed point plays a key role in performance analysis of the supermarket models with repairable servers, it is interesting to develop numerical computation in the study of complex supermarket models.

5.1 A double limit

We discuss a double limit of the expected fraction vector function $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$.

The following lemma provides a sufficient condition under which each of the four interrelated supermarket models with N identical repairable servers is stable.

Lemma 1 *Each of the four supermarket model with N identical and repairable servers and two choice numbers $d_1, d_2 \geq 1$ is stable if $\tilde{\rho} = \rho(1 + \alpha/\beta) < 1$, where $\rho = \lambda/\mu$.*

Proof: If $d_1 = d_2 = 1$, then each of the four supermarket models of N identical repairable servers is equivalent to a system of N independent M/M/1 queues with repairable servers. From Li et al. [34], it is easy to see that such a repairable M/M/1 queue is stable if $\tilde{\rho} < 1$. Using a coupling method, as given in Theorems 4 and 5 of Martin and Suhov [40], it is clear that for a fixed number $N = 1, 2, 3, \dots$, each of the four supermarket models with N identical repairable servers is stable if $\tilde{\rho} < 1$. This completes the proof. ■

The following theorem provides a useful property of the double limit of the expected fraction vector function $\mathbf{u}^{(N)}(t) = (\mathbf{u}_W^{(N)}(t), \mathbf{V}_R^{(N)}(t))$, which is a key to establish the systems of nonlinear equations satisfied by the fixed point.

Theorem 2 *If $\tilde{\rho} = \rho(1 + \alpha/\beta) < 1$, then for each of the four interrelated repairable supermarket models, there exists a unique double limit*

$$\pi = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} \mathbf{u}^{(N)}(t).$$

Proof: This proof is given in Appendix A. ■

In fact, Theorem 2 also gives

$$\pi = \lim_{N \rightarrow \infty} \lim_{t \rightarrow +\infty} \mathbf{u}^{(N)}(t) = \lim_{t \rightarrow +\infty} \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t),$$

which justifies the interchange of the limit of the expected fraction vector function $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$. This is necessary in many practical applications when using the stationary probabilities to give the effective approximation for performance of the supermarket models.

Let $\pi = (\pi_W, \pi_R)$, where $\pi_W = (\pi_{W,0}, \pi_{W,1}, \pi_{W,2}, \dots)$ and $\pi_R = (\pi_{R,1}, \pi_{R,2}, \pi_{R,3}, \dots)$. The row vector π is called a fixed point of the expected fraction vector function $\mathbf{u}^{(N)}(t)$

if $\pi = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} \mathbf{u}^{(N)}(t)$. Based on Theorem 2, we denote by $\pi_{W,k} = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} u_{W,k}^{(N)}(t)$ for $k \geq 0$ and $\pi_{R,l} = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} u_{R,l}^{(N)}(t)$ for $l \geq 1$.

It is well-known that if π is the fixed point of the expected fraction vector function $\mathbf{u}^{(N)}(t)$, then

$$\lim_{t \rightarrow +\infty} \left[\frac{d}{dt} \mathbf{u}^{(N)}(t) \right] = 0,$$

this gives

$$\lim_{t \rightarrow +\infty} \left[\frac{d}{dt} u_{W,k}^{(N)}(t) \right] = 0, k \geq 0; \quad \lim_{t \rightarrow +\infty} \left[\frac{d}{dt} u_{R,l}^{(N)}(t) \right] = 0, l \geq 1.$$

To set up a system of nonlinear equations, we write

$$L_1(\pi_{W,0}, \pi_{W,1}; \pi_{R,1}) = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} L_1^{(N)}(u_{W,0}, u_{W,1}; u_{R,1}; t)$$

and for $k \geq 2$

$$L_k(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} L_k^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t),$$

$$\mathcal{W}_{R,k}^{(N)}(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} \mathcal{W}_{R,k}^{(N)}(u_{W,k-1}, u_{W,k}; u_{R,k-1}, u_{R,k}; t),$$

$$I_1(\pi_{R,1}, \pi_{W,2}) = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} I_1(u_{R,1}^{(N)}(t), u_{W,2}^{(N)}(t))$$

and for $k \geq 2$

$$I_k(\pi_{R,k}, \pi_{W,k+1}) = \lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} I_k(u_{R,k}^{(N)}, u_{W,k+1}^{(N)}; t).$$

It is easy to check from Theorem 1 that

$$(\pi_{W,0} - \pi_{W,1}) L_1(\pi_{W,0}, \pi_{W,1}; \pi_{R,1}) = (\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1}$$

and for $k \geq 2$

$$\begin{aligned} & (\pi_{W,k-1} - \pi_{W,k}) L_k(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) \\ &= \frac{\pi_{W,k-1} - \pi_{W,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right], \end{aligned}$$

$$\begin{aligned} & (\pi_{R,k-1} - \pi_{R,k}) L_k(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) \\ &= \frac{\pi_{R,k-1} - \pi_{R,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right], \end{aligned}$$

$$(\pi_{R,k-1} - \pi_{R,k}) \mathcal{W}_{R,k}^{(N)}(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) = (\pi_{R,k-1} + \pi_{W,k})^{d_1} - (\pi_{R,k} + \pi_{W,k})^{d_1},$$

$$I_1(\pi_{R,1}, \pi_{W,2}) = (1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2}$$

and for $k \geq 2$

$$I_k(\pi_{R,k}, \pi_{W,k+1}) = (1 - \pi_{W,k+1})^{d_2} - (1 - \pi_{R,k} - \pi_{W,k+1})^{d_2}.$$

5.2 Model I ((A.1) and (R.1))

Taking $N \rightarrow \infty$ and $t \rightarrow +\infty$ in both sides of the mean-field equations (27) to (32), it is easy to see that the fixed point satisfies the following system of nonlinear equations

$$-\alpha\pi_{W,1} + \beta\pi_{R,1} = 0, \quad (53)$$

$$\lambda \left[(\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right] - \mu(\pi_{W,1} - \pi_{W,2}) - \alpha\pi_{W,1} + \beta\pi_{R,1} = 0, \quad (54)$$

for $k \geq 2$

$$\begin{aligned} & \lambda \frac{\pi_{W,k-1} - \pi_{W,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\ & - \mu(\pi_{W,k} - \pi_{W,k+1}) - \alpha\pi_{W,k} + \beta\pi_{R,k} = 0, \end{aligned} \quad (55)$$

and

$$\begin{aligned} & \lambda \frac{\pi_{R,k-1} - \pi_{R,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\ & + \alpha\pi_{W,k} - \beta\pi_{R,k} = 0, \end{aligned} \quad (56)$$

with the boundary condition

$$\pi_{W,0} + \pi_{R,1} = 1. \quad (57)$$

To solve the system of nonlinear equations (53) to (57), the following lemma determines the boundary values $\pi_{W,0}$, $\pi_{W,1}$ and $\pi_{R,1}$, which are a key in our computation of the fixed point later.

Lemma 2 *If $\tilde{\rho} < 1$, then*

$$\pi_{W,1} = \frac{\lambda}{\mu} = \rho, \quad (58)$$

$$\pi_{R,1} = \frac{\alpha}{\beta}\rho \quad (59)$$

and

$$\pi_{W,0} = 1 - \frac{\alpha}{\beta}\rho. \quad (60)$$

Proof: It follows from (53) and (57) that

$$\pi_{R,1} = \frac{\alpha}{\beta} \pi_{W,1}$$

and

$$\pi_{W,0} = 1 - \frac{\alpha}{\beta} \pi_{W,1}.$$

It follows from (54) to (56) that

$$\begin{aligned} \pi_{W,1} &= \rho \left\{ (\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right. \\ &\quad \left. + \sum_{k=2}^{\infty} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \right\} \\ &= \rho (\pi_{W,0} + \pi_{R,1})^{d_1} = \rho, \end{aligned}$$

since $\pi_{W,0} + \pi_{R,1} = 1$. This gives

$$\pi_{R,1} = \frac{\alpha}{\beta} \rho$$

and

$$\pi_{W,0} = 1 - \frac{\alpha}{\beta} \rho.$$

This completes the proof. ■

Let $\xi_0 = 1 - \rho\alpha/\beta$, $\xi_1 = \rho$ and $\delta_1 = \rho\alpha/\beta$. Using (53) and (57), we take that $\pi_{W,0} = \xi_0$, $\pi_{W,1} = \xi_1$ and $\pi_{R,1} = \delta_1$. Let

$$\xi_2 = \xi_1 - \rho (\xi_0 - \xi_1) L_1 (\xi_0, \xi_1; \delta_1) + \frac{\alpha}{\mu} \xi_1 - \frac{\beta}{\mu} \delta_1 \quad (61)$$

and δ_2 the unique solution in $(0, \delta_1)$ to the nonlinear equation

$$F_2(x) = \beta x - \lambda (\delta_1 - x) L_2 (\xi_1, \xi_2; \delta_1, x) - \alpha \xi_2 = 0. \quad (62)$$

For $l \geq 3$, we set

$$F_l(x) = \beta x - \lambda (\delta_{l-1} - x) L_l (\xi_{l-1}, \xi_l; \delta_{l-1}, x) - \alpha \xi_l. \quad (63)$$

We assume that for $l \leq k-1$, the $k-1$ pairs $(\xi_0, \delta_1), (\xi_1, \delta_2), \dots, (\xi_{k-2}, \delta_{k-1})$ have been given iteratively, where δ_{k-1} is the unique solution in $(0, \delta_{k-2})$ to the nonlinear equation $F_{k-1}(x) = 0$. For $l = k$, we write

$$\xi_k = \xi_{k-1} - \rho (\xi_{k-2} - \xi_{k-1}) L_{k-1} (\xi_{k-2}, \xi_{k-1}; \delta_{k-2}, \delta_{k-1}) + \frac{\alpha}{\mu} \xi_{k-1} - \frac{\beta}{\mu} \delta_{k-1}, \quad (64)$$

and δ_k is the unique solution in $(0, \delta_{k-1})$ to the nonlinear equation $F_k(x) = 0$. It is clear that $0 < \xi_k < \xi_{k-1} < \dots < \xi_1 < \xi_0 = 1 - \rho\alpha/\beta$ and $0 < \delta_k < \delta_{k-1} < \dots < \delta_2 < \delta_1 = \rho\alpha/\beta$.

The following theorem provides expression for the fixed point by means of the system of nonlinear equations (53) to (57).

Theorem 3 *If $\tilde{\rho} < 1$, then the fixed point $\pi = (\pi_{W,0}, \pi_{W,1}, \pi_{W,2}, \dots; \pi_{R,1}, \pi_{R,2}, \pi_{R,3}, \dots)$ is given by*

$$\pi_{W,k} = \xi_k, k \geq 0,$$

and

$$\pi_{R,l} = \delta_l, l \geq 1.$$

Proof: Lemma 2 shows that $\pi_{W,0} = \xi_0$, $\pi_{W,1} = \xi_1$ and $\pi_{R,1} = \delta_1$.

We assume that for $1 \leq l \leq k$, $\pi_{W,l} = \xi_l$ and $\pi_{R,l} = \delta_l$, where $0 < \xi_k < \xi_{k-1} < \dots < \xi_1 < \xi_0 = 1 - \rho\alpha/\beta$ and $0 < \delta_k < \delta_{k-1} < \dots < \delta_2 < \delta_1 = \rho\alpha/\beta$. Then for $l = k+1$, it follows from Equation (54) that

$$\begin{aligned} \pi_{W,k+1} &= \pi_{W,k} - \rho(\pi_{W,k-1} - \pi_{W,k}) L_k(\pi_{W,k-1}, \pi_{W,k}; \pi_{R,k-1}, \pi_{R,k}) + \frac{\alpha}{\mu} \pi_{W,k} - \frac{\beta}{\mu} \pi_{R,k} \\ &= \xi_k - \rho(\xi_{k-1} - \xi_k) L_k(\xi_{k-1}, \xi_k; \delta_{k-1}, \delta_k) + \frac{\alpha}{\mu} \xi_k - \frac{\beta}{\mu} \delta_k = \xi_{k+1}. \end{aligned}$$

It follows from Equation (56) that

$$\lambda(\delta_k - \pi_{R,k+1}) L_{k+1}(\xi_k, \xi_{k+1}; \delta_k, \pi_{R,k+1}) + \alpha\xi_{k+1} - \beta\pi_{R,k+1} = 0.$$

Let

$$\begin{aligned} F_{k+1}(x) &= \beta x - \lambda(\delta_k - x) L_{k+1}(\xi_k, \xi_{k+1}; \delta_k, x) - \alpha\xi_{k+1} \\ &= \beta x - \frac{\lambda(\delta_k - x)}{\xi_k - \xi_{k+1} + \delta_k - x} \left[(\xi_k + \delta_k)^{d_1} - (\xi_{k+1} + x)^{d_1} \right] - \alpha\xi_{k+1}. \end{aligned}$$

Then

$$\begin{aligned} F_{k+1}(0) &= -\frac{\lambda\delta_k}{\xi_k - \xi_{k+1} + \delta_k} \left[(\xi_k + \delta_k)^{d_1} - \xi_{k+1}^{d_1} \right] - \alpha\xi_k < 0, \\ F_{k+1}(\delta_k) &= \beta\delta_k - \alpha\xi_{k+1} > \beta\delta_k - \alpha\xi_k = \lambda(\delta_{k-1} - \delta_k) L_k(\xi_{k-1}, \xi_k; \delta_{k-1}, \delta_k) > 0 \end{aligned}$$

by means of (56), and

$$\begin{aligned}
\frac{d}{dx}F_{k+1}(x) &= \beta - \frac{d}{dx} \left[\frac{\lambda(\delta_k - x)(\xi_k + \delta_k)^{d_1}}{\xi_k - \xi_{k+1} + \delta_k - x} \right] + \frac{d}{dx} \left[\frac{\lambda(\delta_k - x)(\xi_{k+1} + x)^{d_1}}{\xi_k - \xi_{k+1} + \delta_k - x} \right] \\
&= \beta + \lambda \frac{(\xi_k + \delta_k)^{d_1}(\xi_k - \xi_{k+1})}{(\xi_k - \xi_{k+1} + \delta_k - x)^2} - \lambda \frac{(\xi_{k+1} + x)^{d_1}(\xi_k - \xi_{k+1})}{(\xi_k - \xi_{k+1} + \delta_k - x)^2} \\
&\quad + \lambda \frac{d_1(\xi_{k+1} + x)^{d_1-1}(\delta_k - x)(\xi_k - \xi_{k+1} + \delta_k - x)}{(\xi_k - \xi_{k+1} + \delta_k - x)^2} \\
&\geq \beta + \lambda \frac{d_1(\xi_{k+1} + x)^{d_1-1}(\delta_k - x)(\xi_k - \xi_{k+1} + \delta_k - x)}{(\xi_k - \xi_{k+1} + \delta_k - x)^2} > 0
\end{aligned}$$

by means of $\xi_k + \delta_k \geq \xi_{k+1} + x$. Note that $F_{k+1}(x)$ is a continuous function for $x \in (0, \delta_k)$, there exists a unique positive solution δ_{k+1} in $(0, \delta_k)$ to the nonlinear equation $F_{k+1}(x) = 0$. Hence, $\pi_{R,k+1} = \delta_{k+1}$.

By induction, this completes the proof. ■

Note that for the other three models with more complex complex nonlinear structures, we provide some discussion on the boundary conditions: $\pi_{W,0}, \pi_{W,1}$ and $\pi_{R,1}$.

5.3 Model II ((A.1) and (R.2))

Taking $N \rightarrow \infty$ and $t \rightarrow +\infty$ in both sides of the mean-field equations (33) to (37), it is easy to see that the fixed point satisfies the following system of nonlinear equations

$$-\alpha\pi_{W,1} + \beta \left[(1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2} \right] = 0, \quad (65)$$

$$\begin{aligned}
&\lambda \left[(\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right] - \mu(\pi_{W,1} - \pi_{W,2}) - \alpha\pi_{W,1} \\
&+ \beta \left[(1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2} \right] = 0,
\end{aligned} \quad (66)$$

for $k \geq 2$

$$\begin{aligned}
&\lambda \frac{\pi_{W,k-1} - \pi_{W,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\
&- \mu(\pi_{W,k} - \pi_{W,k+1}) - \alpha\pi_{W,k} + \beta \left[(1 - \pi_{W,k+1})^{d_2} - (1 - \pi_{R,k} - \pi_{W,k+1})^{d_2} \right] = 0,
\end{aligned} \quad (67)$$

and

$$\begin{aligned}
&\lambda \frac{\pi_{R,k-1} - \pi_{R,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\
&+ \alpha\pi_{W,k} - \beta \left[(1 - \pi_{W,k+1})^{d_2} - (1 - \pi_{R,k} - \pi_{W,k+1})^{d_2} \right] = 0,
\end{aligned} \quad (68)$$

with the boundary condition

$$\pi_{W,0} + \pi_{R,1} = 1. \quad (69)$$

It follows from (65) and (66) that

$$\lambda \left[(\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right] - \mu (\pi_{W,1} - \pi_{W,2}) = 0,$$

and from (67) and (68) that for $k \geq 2$

$$\lambda \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] - \mu (\pi_{W,k} - \pi_{W,k+1}) = 0,$$

which, together with (69), follows

$$\pi_{W,1} = \rho (\pi_{W,0} + \pi_{R,1})^{d_1} = \rho. \quad (70)$$

It follows from (65), (66) and (70) that

$$\pi_{W,2} = \rho (\rho + \pi_{R,1})^{d_1}. \quad (71)$$

From (70), (71) and (66), we find that $\pi_{R,1}$ is the minimal nonnegative solution to the following nonlinear equation

$$\left[1 - \rho (\rho + \pi_{R,1})^{d_1} \right]^{d_2} - \left[1 - \pi_{R,1} - \rho (\rho + \pi_{R,1})^{d_1} \right]^{d_2} = \rho \frac{\alpha}{\beta}.$$

Also, $\pi_{W,0} = 1 - \pi_{R,1}$ is given.

5.4 Model III ((A.2) and (R.1))

Taking $N \rightarrow \infty$ and $t \rightarrow +\infty$ in both sides of the mean-field equations (41) to (45), we obtain that the fixed point satisfies the following system of nonlinear equations

$$-\alpha \pi_{W,1} + \beta \pi_{R,1} = 0, \quad (72)$$

$$\lambda \left[(\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right] - \mu (\pi_{W,1} - \pi_{W,2}) - \alpha \pi_{W,1} + \beta \pi_{R,1} = 0, \quad (73)$$

for $k \geq 2$

$$\begin{aligned} & \lambda \frac{\pi_{W,k-1} - \pi_{W,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\ & - \mu (\pi_{W,k} - \pi_{W,k+1}) - \alpha \pi_{W,k} + \beta \pi_{R,k} = 0, \end{aligned} \quad (74)$$

and

$$\lambda \left[(\pi_{R,k-1} + \pi_{W,k})^{d_1} - (\pi_{R,k} + \pi_{W,k})^{d_1} \right] + \alpha \pi_{W,k} - \beta \pi_{R,k} = 0, \quad (75)$$

with the boundary condition

$$\pi_{W,0} + \pi_{R,1} = 1. \quad (76)$$

Now, we discuss the boundary conditions of the fixed point. It follows from (72) and (76) that

$$\pi_{R,1} = \frac{\alpha}{\beta} \pi_{W,1}$$

and

$$\pi_{W,0} = 1 - \frac{\alpha}{\beta} \pi_{W,1}.$$

It follows from (73) to (75) that

$$\begin{aligned} \frac{1}{\rho} \pi_{W,1} &= 1 + \sum_{k=2}^{\infty} \left[(\pi_{R,k-1} + \pi_{W,k})^{d_1} - (\pi_{R,k} + \pi_{W,k})^{d_1} \right] \\ &\quad - \sum_{k=2}^{\infty} \frac{\pi_{R,k-1} - \pi_{R,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{R,k-1} + \pi_{W,k-1})^{d_1} - (\pi_{R,k} + \pi_{W,k})^{d_1} \right] \\ &= 1 + \sum_{k=2}^{\infty} \left[(\pi_{W,k-1} + \pi_{R,k})^{d_1} \right. \\ &\quad \left. - \frac{(\pi_{W,k-1} - \pi_{W,k})(\pi_{W,k} + \pi_{R,k})^{d_1} + (\pi_{R,k-1} - \pi_{R,k})(\pi_{W,k-1} + \pi_{R,k-1})^{d_1}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \right]. \end{aligned}$$

Let

$$\begin{aligned} \Delta(d_1) &= \sum_{k=2}^{\infty} \frac{(\pi_{W,k-1} - \pi_{W,k})(\pi_{W,k} + \pi_{R,k})^{d_1} + (\pi_{R,k-1} - \pi_{R,k})(\pi_{W,k-1} + \pi_{R,k-1})^{d_1}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \\ &\quad - \sum_{k=2}^{\infty} (\pi_{W,k-1} + \pi_{R,k})^{d_1}. \end{aligned}$$

This gives

$$\pi_{W,1} = \rho [1 - \Delta(d_1)].$$

It is easy to check that $\Delta(d_1) \in (0, 1)$, and $\Delta(d_1)$ is increasing in $d_1 \geq 1$. Therefore $\pi_{W,1} < \rho$ if $d_1 \geq 2$, and $\pi_{W,1} = \rho$ if $d_1 = 1$. At the same time, $\pi_{W,1}$ is decreasing in $d_1 \geq 1$.

5.5 Model IV ((A.2) and (R.2))

Taking $N \rightarrow \infty$ and $t \rightarrow +\infty$ in both sides of the mean-field equations (47) to (51), it is easy to see that the fixed point satisfies the following system of nonlinear equations

$$-\alpha \pi_{W,1} + \beta \left[(1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2} \right] = 0, \quad (77)$$

$$\begin{aligned} & \lambda \left[(\pi_{W,0} + \pi_{R,1})^{d_1} - (\pi_{W,1} + \pi_{R,1})^{d_1} \right] - \mu (\pi_{W,1} - \pi_{W,2}) - \alpha \pi_{W,1} \\ & + \beta \left[(1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2} \right] = 0, \end{aligned} \quad (78)$$

for $k \geq 2$

$$\begin{aligned} & \lambda \frac{\pi_{W,k-1} - \pi_{W,k}}{\pi_{W,k-1} - \pi_{W,k} + \pi_{R,k-1} - \pi_{R,k}} \left[(\pi_{W,k-1} + \pi_{R,k-1})^{d_1} - (\pi_{W,k} + \pi_{R,k})^{d_1} \right] \\ & - \mu (\pi_{W,k} - \pi_{W,k+1}) - \alpha \pi_{W,k} + \beta \left[(1 - \pi_{W,k+1})^{d_2} - (1 - \pi_{R,k} - \pi_{W,k+1})^{d_2} \right] = 0, \end{aligned} \quad (79)$$

and

$$\begin{aligned} & \lambda \left[(\pi_{R,k-1} + \pi_{W,k})^{d_1} - (\pi_{R,k} + \pi_{W,k})^{d_1} \right] \\ & + \alpha \pi_{W,k} - \beta \left[(1 - \pi_{W,k+1})^{d_2} - (1 - \pi_{R,k} - \pi_{W,k+1})^{d_2} \right] = 0, \end{aligned} \quad (80)$$

with the boundary condition

$$\pi_{W,0} + \pi_{R,1} = 1. \quad (81)$$

From the similar analysis to the boundary conditions in Model III, we obtain

$$\pi_{W,1} = \rho [1 - \Delta(d_1)],$$

and the similar analysis to that in Model II leads to

$$\pi_{W,2} = -\rho \Delta(d_1) + \rho \{ \rho [1 - \Delta(d_1)] + \pi_{R,1} \}^{d_1},$$

and $\pi_{R,1}$ is the minimal nonnegative solution to the following nonlinear equation

$$(1 - \pi_{W,2})^{d_2} - (1 - \pi_{R,1} - \pi_{W,2})^{d_2} = \pi_{W,1} \frac{\alpha}{\beta}.$$

Also, we get that $\pi_{W,0} = 1 - \pi_{R,1}$.

6 Performance Analysis and Numerical Observations

In this section, we first provide useful performance measures of the four interrelated supermarket models with repairable servers. Then we use some numerical examples to make valuable observations on model improvement by means of performance numerical comparison.

6.1 Performance measures

(a) The mean of stationary queueing length

Let \mathcal{Q} be the stationary queue length of any server in each of the four supermarket models. Then

$$\begin{aligned} E[\mathcal{Q}] &= \sum_{k=1}^{\infty} P\{\mathcal{Q} \geq k\} \\ &= \sum_{k=1}^{\infty} \{P\{\mathcal{Q} \geq k, \text{ the server is working}\} + P\{\mathcal{Q} \geq k, \text{ the server is repairing}\}\} \\ &= \sum_{k=1}^{\infty} (\pi_{W,k} + \pi_{R,k}). \end{aligned}$$

(b) The variance of stationary queueing length

It is easy to check that

$$Var[\mathcal{Q}] = \sum_{k=1}^{\infty} (2k-1) (\pi_{W,k} + \pi_{R,k}) - \left[\sum_{k=1}^{\infty} (\pi_{W,k} + \pi_{R,k}) \right]^2,$$

since

$$\begin{aligned} E[\mathcal{Q}^2] &= \sum_{k=1}^{\infty} k^2 [P\{\mathcal{Q} \geq k\} - P\{\mathcal{Q} \geq k+1\}] \\ &= \sum_{k=1}^{\infty} (2k-1) P\{\mathcal{Q} \geq k\}. \end{aligned}$$

(c) The steady-state availability and failure frequency

Let A and W_f be the steady-state availability and failure frequency in any repairable server, respectively. Then

$$A = \sum_{k=0}^{\infty} (\pi_{W,k} - \pi_{W,k+1}) = \pi_{W,0} \quad (82)$$

and

$$W_f = \alpha \sum_{k=1}^{\infty} (\pi_{W,k} - \pi_{W,k+1}) = \alpha \pi_{W,1}. \quad (83)$$

(d) The steady-state mean-field flow balancing

Since the mean-field theory plays an important role in the study of supermarket models, a flow balancing in the supermarket models is called a mean-field flow balancing. In every supermarket model, the steady-state mean-field throughput is given by

$$\text{MF-TH} = \mu \pi_{W,1}.$$

Thus the the steady-state mean-field input-output difference is defined as

$$F(d_1, d_2) = \lambda - \mu\pi_{W,1}.$$

Clearly, this supermarket model can have a steady-state mean-field flow balancing if $F(d_1, d_2) = 0$.

6.2 Numerical observations

Now, we use some numerical examples to show how the major performance measures depend on some crucial parameters of the systems. These numerical examples are organized in three groups for different purposes: (1) examining the mean $E[\mathcal{Q}]$ and the variance $Var[\mathcal{Q}]$; (2) observing the availability A and the failure frequency W_f ; and (3) discussing the mean-field flow balancing $F(d_1, d_2)$. At the same time, it is worth noting that in the numerical examples, (A.1) and (A.2) represent routing customers and (R.1) and (R.2) represent organizing of repair resource.

The following seven numerical examples are based on a set of system parameters of $\mu = 9$, $\alpha = 2$, $\beta = 5$ and $\lambda \in (0, 6)$. It is easy to check that $\tilde{\rho} < 42/45 < 1$.

Example 1: Show $E[\mathcal{Q}]$ in Models III and IV for a comparison of deployment of repair resource. In this example with (A.2), we consider Models III and IV and observe how choice numbers d_1 and d_2 affect $E[\mathcal{Q}]$, the stationary queue length. Figure 3 shows how $E[\mathcal{Q}]$ changes with $d_1, d_2 = 1, 2, 3$ when the arrival rate $\lambda \in (0, 6)$. It is observed that while $E[\mathcal{Q}]$ increases with λ , it decreases with either d_1 or d_2 . Also, d_1 is more effective than d_2 in terms of reducing $E[\mathcal{Q}]$.

Example 2: Focus on $E[\mathcal{Q}]$ to compare (A.1) with (A.2)

In this example, we demonstrate how (A.2) improves the performance under (A.1) in terms of $E[\mathcal{Q}]$. Figure 4 shows the $E[\mathcal{Q}]$ as a function of the arrival rate $\lambda \in (0, 6)$ with $d_1, d_2 = 2, 3$. It is observed that (A.2) can effectively reduce $E[\mathcal{Q}]$ compared with (A.1). This implies that using more system information can improve the system performance.

Example 3: Show $Var[\mathcal{Q}]$ in Models III and IV for comparing repair resource deployment.

In this example with (A.2), we focus on how d_1 and d_2 affect $Var[\mathcal{Q}]$, the variance of stationary queue length. Figure 5 illustrates how $Var[\mathcal{Q}]$ changes with the arrival rate $\lambda \in (0, 6)$ with $d_1, d_2 = 1, 2, 3$. We observe that the mean queue length decreases with either d_1 or d_2 . Also, d_1 is more effective than d_2 in terms of reducing $Var[\mathcal{Q}]$.

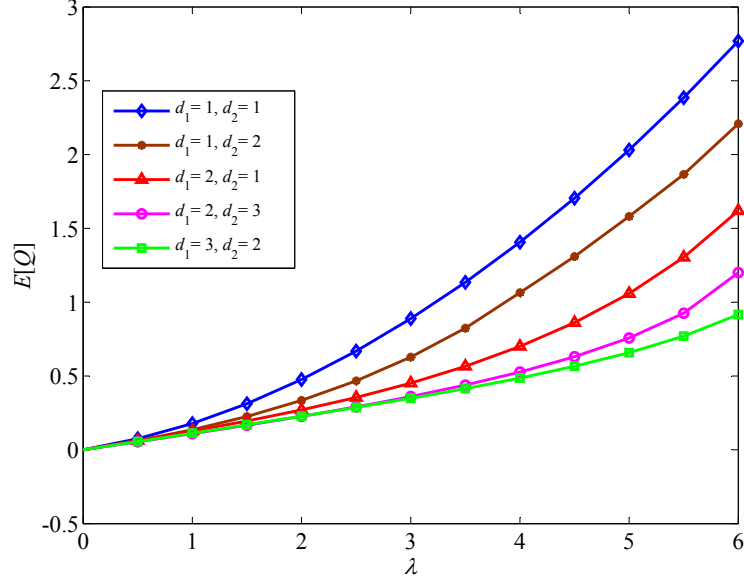


Figure 3: $E[Q]$ for a comparison of repair organizations

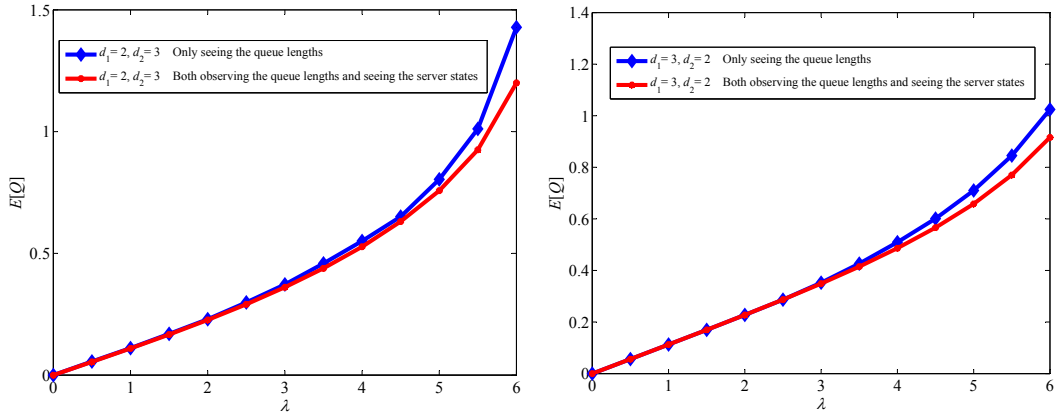


Figure 4: $E[Q]$ for comparing (A.1) with (A.2)

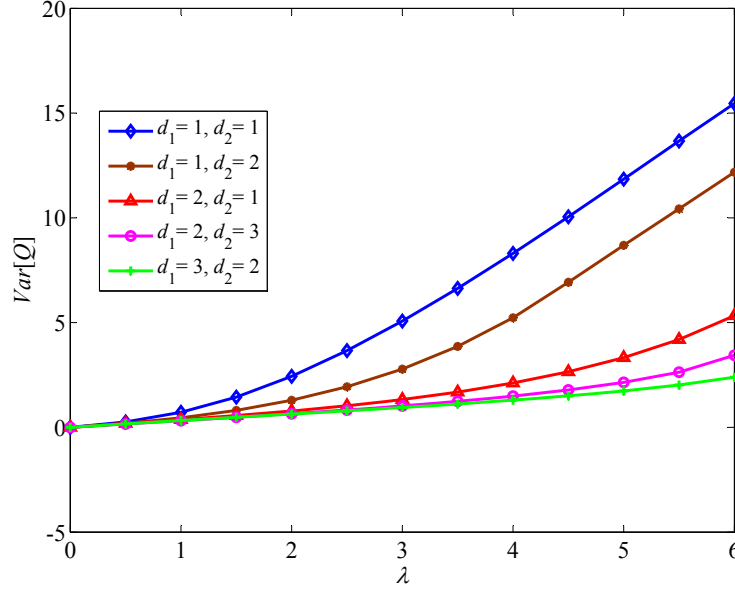


Figure 5: $Var[Q]$ for a comparison of repair organizations

Example 4: Observe $Var[Q]$ under (A.1) or (A.2)

In this example, Figure 6 shows how $Var[Q]$ changes on the arrival rate $\lambda \in (0, 6)$ with $d_1, d_2 = 2, 3$. It is revealed that (A.2) can effectively reduce $Var[Q]$ under (A.1).

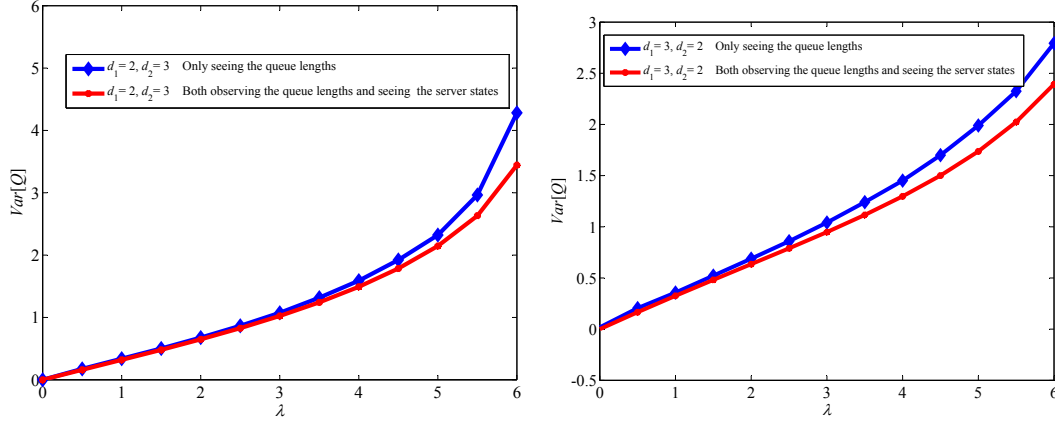


Figure 6: $Var[Q]$ for comparing (A.1) with (A.2)

Example 5: Examine the steady-state availability A in Models III and IV

In this example with (A.2), Figure 7 shows that while the steady-state availability A decreases with λ , it increases with either d_1 or d_2 . Thus, d_1 and d_2 can help increase the steady-state availability.

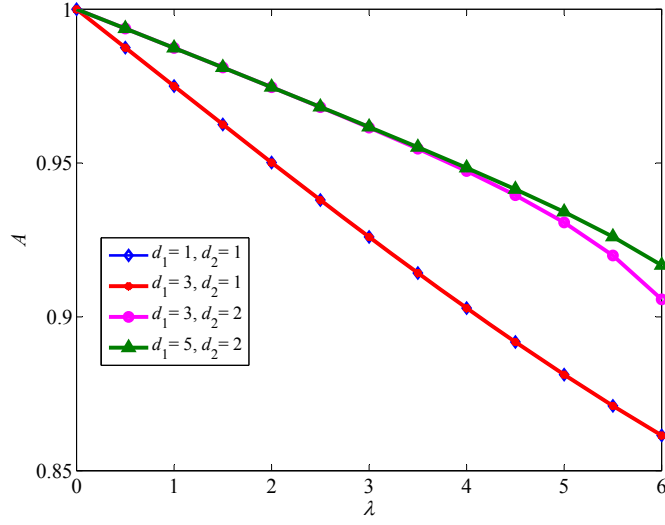


Figure 7: A in Models III and IV

Example 6: Investigate the steady-state failure frequency W_f in Models III and IV

In this example with (A.2), Figure 8 shows that W_f increases with both λ and d_1 or d_2 .

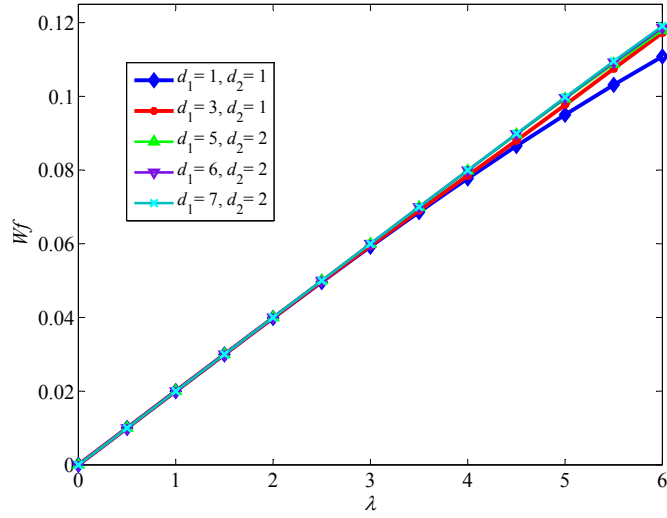


Figure 8: W_f in Models III and IV

Finally, we provide a numerical example to show the steady-state flow balancing in the study of supermarket models.

Example 7: Observe the steady-state mean-field flow balancing

In this example with (A.2), we show how the steady-state mean-field input-output difference $F(d_1, d_2)$ depends on the arrival rate $\lambda \in (0, 5)$ with $d_1 = 1, 5, 6$ and $d_2 = 1, 2, 3$.

Figure 9 indicates that if $d_1 = 5, 6$ and $d_2 = 2, 3$, the steady-state mean-field input-output difference $F(d_1, d_2) > 0$, and it increases with λ . However, $F(1, 1) = 0$, which implies that the repairable supermarket model has the steady-state mean-field flow balancing for $d_1, d_2 = 1$.

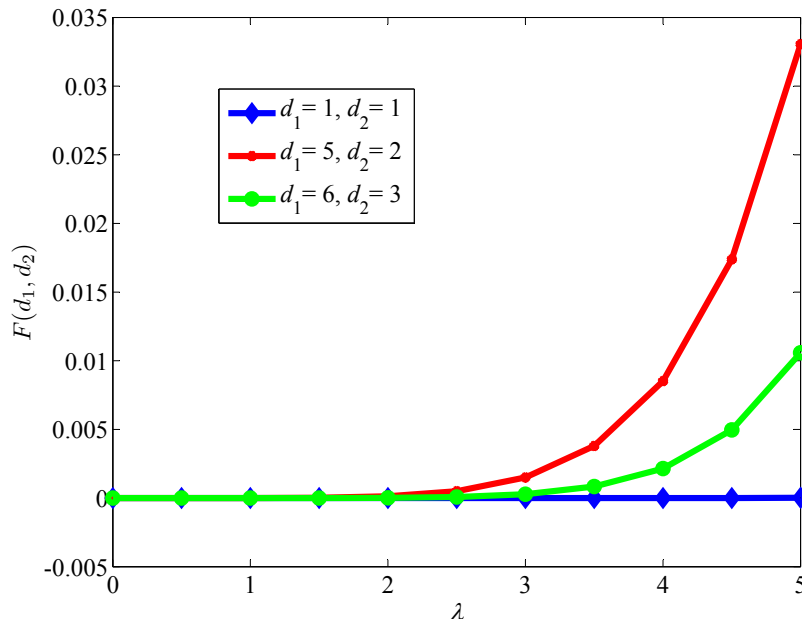


Figure 9: $F(d_1, d_2)$ in Models III and IV

From the numerical analysis above, we may conclude that the system information (i.e., server in working or repair condition and queue length) for the arriving customer and the deployment of the repair resource can effectively improve the system performances of the supermarket models.

7 Concluding Remarks

In this paper, we apply the mean-field theory to studying effects of a double dynamic routine selection scheme (for the arrival dispatched schemes and for the groups of repairmen) on performance of the four interrelated supermarket models with repairable servers. We first provide a probability method of setting up the infinite-dimensional systems of mean-field equations. Then we prove asymptotic independence of the supermarket mod-

els with repairable servers. Based on this, we discuss the fixed points which are computed by means of the systems of nonlinear equations. Finally, we provide useful performance measures of the supermarket models, and use some numerical examples to make valuable observations on model improvement via using system information and deploying repair resource. Our results reveal effects of utilizing system information for customer's joining decisions as well as reorganization of repair resource on performance of the supermarket models. Along with this line, there are a number of interesting directions for future research, for example:

- analyzing non-Poisson inputs, such as, Markovian arrival processes (MAPs), and renewal processes;
- studying non-exponential service time distributions, for example, general distributions, matrix-exponential distributions and heavy-tailed distributions;
- discussing the bulk arrival processes, and the bulk service processes; and
- developing effective algorithms for computing the fixed points in the study of complex supermarket models.

Acknowledgements

The first two authors were supported by the National Natural Science Foundation of China under grant No. 71471160, No. 71671158 and No. 71471114, and the Fostering Plan of Innovation Team and Leading Talent in Hebei Universities under grant No. LJRC027.

Appendix A: The Proof of Theorem 2

In this appendix, for the four interrelated supermarket models with repairable servers, we provide a simple outline of the proof of Theorem 2. To that end, it is a key to use the operator semigroup to provide a mean-field limit for the sequence of Markov processes who asymptotically approaches a single trajectory identified by the unique and global solution to an infinite-dimensional system of mean-field equations. Readers may refer to Li et al. [30] for more details with respect to the proof of such a mean-field limit.

For the vector $\mathbf{u}^{(N)} = (\mathbf{u}_W^{(N)}, \mathbf{u}_R^{(N)})$ where $\mathbf{u}_W^{(N)} = (u_0^{(N)}, u_1^{(N)}, u_2^{(N)}, \dots)$ and $\mathbf{u}_R^{(N)} = (v_1^{(N)}, v_2^{(N)}, v_3^{(N)}, \dots)$, we write

$$\begin{aligned} \tilde{\Omega}_N = \Big\{ \mathbf{u}^{(N)} = (\mathbf{u}_W^{(N)}, \mathbf{u}_R^{(N)}) : & 1 \geq u_0^{(N)} \geq u_1^{(N)} \geq u_2^{(N)} \geq u_3^{(N)} \geq \dots \geq 0, \\ & 1 \geq v_1^{(N)} \geq v_2^{(N)} \geq v_3^{(N)} \geq v_4^{(N)} \geq \dots \geq 0, \\ & Nu_k^{(N)} \text{ and } Nv_l^{(N)} \text{ are nonnegative integers for } k \geq 0 \text{ and } l \geq 1 \Big\} \end{aligned}$$

and

$$\Omega_N = \left\{ \mathbf{u}^{(N)} \in \tilde{\Omega}_N : \mathbf{u}^{(N)} e < +\infty \right\}.$$

At the same time, for the vector $\mathbf{u} = (\mathbf{u}_W, \mathbf{u}_R)$ where $\mathbf{u}_W = (u_0, u_1, u_2, \dots)$ and $\mathbf{u}_R = (v_1, v_2, v_3, \dots)$, we set

$$\tilde{\Omega} = \{ \mathbf{u} = (\mathbf{u}_W, \mathbf{u}_R) : 1 \geq u_0 \geq u_1 \geq \dots \geq 0, 1 \geq v_1 \geq v_2 \geq \dots \}$$

and

$$\Omega = \left\{ \mathbf{u} \in \tilde{\Omega} : \mathbf{u} e < +\infty \right\}.$$

Obviously, $\Omega_N \subsetneq \Omega \subsetneq \tilde{\Omega}$ and $\Omega_N \subsetneq \tilde{\Omega}_N \subsetneq \tilde{\Omega}$.

In the infinite-dimensional vector space $\tilde{\Omega}$, we take a metric

$$\rho(\mathbf{u}, \mathbf{u}') = \sup_{k \geq 0, l \geq 1} \left\{ \frac{|u_k - u'_k|}{k+1}, \frac{|v_l - v'_l|}{l} \right\}, \quad (84)$$

for $\mathbf{u}, \mathbf{u}' \in \tilde{\Omega}$. Note that under the metric $\rho(\mathbf{u}, \mathbf{u}')$, the infinite-dimensional vector space $\tilde{\Omega}$ is complete, separable and compact.

For simplicity of description, here we only study the sequence $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes in the first supermarket model with repairable servers, while the other three models can be analyzed similarly without any difficulty.

For the first supermarket model with repairable servers, the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ is described as

$$\frac{d}{dt} \mathbf{U}^{(N)}(t) = \mathbf{A}_N f(\mathbf{U}^{(N)}(t)),$$

where \mathbf{A}_N acting on functions $f : \Omega_N \rightarrow \mathbf{C}^1$ is the generating operator of the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$, and

$$\mathbf{A}_N = \mathbf{A}_N^{\text{Input}} + \mathbf{A}_N^{\text{Out}} + \mathbf{A}_N^{\text{Transition}}, \quad (85)$$

for $\mathbf{u} = (\mathbf{g}, \mathbf{h})$, $\mathbf{g} = (g_0, g_1, g_2, \dots)$ and $\mathbf{h} = (h_1, h_2, h_3, \dots)$,

$$\begin{aligned} \mathbf{A}_N^{\text{Input}} &= \lambda N (g_0 - g_1) L_1 (g_0, g_1; h_1) \left[f(\mathbf{g} + \frac{\mathbf{e}_1}{N}, \mathbf{h}) - f(\mathbf{g}, \mathbf{h}) \right] \\ &\quad + \lambda N \sum_{k=2}^{\infty} \left\{ (g_{k-1} - g_k) L_k (g_{k-1}, g_k; h_{k-1}, h_k) \left[f(\mathbf{g} + \frac{\mathbf{e}_k}{N}, \mathbf{h}) - f(\mathbf{g}, \mathbf{h}) \right] \right. \\ &\quad \left. + (h_{k-1} - h_k) L_k (g_{k-1}, g_k; h_{k-1}, h_k) \left[f(\mathbf{g}, \mathbf{h} + \frac{\mathbf{e}_k}{N}) - f(\mathbf{g}, \mathbf{h}) \right] \right\}, \\ \mathbf{A}_N^{\text{Out}} &= \mu N \sum_{k=1}^{\infty} (g_k - g_{k+1}) \left[f\left(\mathbf{g} - \frac{\mathbf{e}_k}{N}, \mathbf{h}\right) - f(\mathbf{g}, \mathbf{h}) \right] \end{aligned}$$

and

$$\begin{aligned} \mathbf{A}_N^{\text{Transition}} &= \alpha N \sum_{k=1}^{\infty} g_k \left[f\left(\mathbf{g} - \frac{\mathbf{e}_k}{N}, \mathbf{h} + \frac{\mathbf{e}_k}{N}\right) - f(\mathbf{g}, \mathbf{h}) \right] \\ &\quad + \beta N \sum_{k=1}^{\infty} h_k \left[f\left(\mathbf{g} + \frac{\mathbf{e}_k}{N}, \mathbf{h} - \frac{\mathbf{e}_k}{N}\right) - f(\mathbf{g}, \mathbf{h}) \right], \end{aligned}$$

where \mathbf{e}_k stands for a row vector with the k th entry be one and all the other entries be zero, and

$$L_1 (g_0; g_1, h_1) = \sum_{m=1}^{d_1} C_{d_1}^m (g_0 - g_1)^{m-1} (g_1 + h_1)^{d-m},$$

for $k \geq 2$

$$L_k (g_{k-1}, g_k; h_{k-1}, h_k) = \sum_{m=1}^{d_1} C_{d_1}^m (g_{k-1} - g_k + h_{k-1} - h_k)^{m-1} (g_k + h_k)^{d-m}.$$

Therefore, for $\mathbf{u} = (\mathbf{g}, \mathbf{h}) \in \Omega_N$ and the function $f : \Omega_N \rightarrow \mathbf{C}^1$ we obtain

$$\begin{aligned} \mathbf{A}_N f(\mathbf{g}, \mathbf{h}) &= \lambda N (g_0 - g_1) L_1 (g_0, g_1; h_1) \left[f(\mathbf{g} + \frac{\mathbf{e}_1}{N}, \mathbf{h}) - f(\mathbf{g}, \mathbf{h}) \right] \\ &\quad + \lambda N \sum_{k=2}^{\infty} \left\{ (g_{k-1} - g_k) L_k (g_{k-1}, g_k; h_{k-1}, h_k) \left[f(\mathbf{g} + \frac{\mathbf{e}_k}{N}, \mathbf{h}) - f(\mathbf{g}, \mathbf{h}) \right] \right. \\ &\quad \left. + (h_{k-1} - h_k) L_k (g_{k-1}, g_k; h_{k-1}, h_k) \left[f(\mathbf{g}, \mathbf{h} + \frac{\mathbf{e}_k}{N}) - f(\mathbf{g}, \mathbf{h}) \right] \right\} \\ &\quad + \mu N \sum_{k=1}^{\infty} (g_k - g_{k+1}) \left[f\left(\mathbf{g} - \frac{\mathbf{e}_k}{N}, \mathbf{h}\right) - f(\mathbf{g}, \mathbf{h}) \right] \\ &\quad + \alpha N \sum_{k=1}^{\infty} g_k \left[f\left(\mathbf{g} - \frac{\mathbf{e}_k}{N}, \mathbf{h} + \frac{\mathbf{e}_k}{N}\right) - f(\mathbf{g}, \mathbf{h}) \right] \\ &\quad + \beta N \sum_{k=1}^{\infty} h_k \left[f\left(\mathbf{g} + \frac{\mathbf{e}_k}{N}, \mathbf{h} - \frac{\mathbf{e}_k}{N}\right) - f(\mathbf{g}, \mathbf{h}) \right]. \end{aligned} \tag{86}$$

The operator semigroup of the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ is defined as $\mathbf{T}_N(t)$, where if $f : \Omega_N \rightarrow \mathbf{C}^1$, then for $(\mathbf{g}, \mathbf{h}) \in \Omega_N$ and $t \geq 0$

$$\mathbf{T}_N(t)f(\mathbf{g}, \mathbf{h}) = E[f(\mathbf{U}_N(t) \mid \mathbf{U}_N(0) = (\mathbf{g}, \mathbf{h})].$$

Note that \mathbf{A}_N is the generating operator of the operator semigroup $\mathbf{T}_N(t)$, it is easy to see that $\mathbf{T}_N(t) = \exp\{\mathbf{A}_N t\}$ for $t \geq 0$.

To analyze the limiting behavior of the sequence $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ of the Markov processes, two formal limits for the sequence $\{\mathbf{A}_N\}$ of the generating operators and for the sequence $\{\mathbf{T}_N(t)\}$ of the semigroups are expressed as $\mathbf{A} = \lim_{N \rightarrow \infty} \mathbf{A}_N$ and $\mathbf{T}(t) = \lim_{N \rightarrow \infty} \mathbf{T}_N(t)$ for $t \geq 0$, respectively. It follows from (86) that as $N \rightarrow \infty$

$$\begin{aligned} \mathbf{A}f(\mathbf{g}, \mathbf{h}) = & \lambda N (g_0 - g_1) L_1(g_0, g_1; h_1) \frac{\partial}{\partial g_1} f(\mathbf{g}, \mathbf{h}) \\ & + \lambda N \sum_{k=2}^{\infty} \left[(g_{k-1} - g_k) L_k(g_{k-1}, g_k; h_{k-1}, h_k) \frac{\partial}{\partial g_k} f(\mathbf{g}, \mathbf{h}) \right. \\ & \left. + (h_{k-1} - h_k) L_k(g_{k-1}, g_k; h_{k-1}, h_k) \frac{\partial}{\partial h_k} f(\mathbf{g}, \mathbf{h}) \right] \\ & - \mu N \sum_{k=1}^{\infty} (g_k - g_{k+1}) \frac{\partial}{\partial g_k} f(\mathbf{g}, \mathbf{h}) \\ & - \alpha N \sum_{k=1}^{\infty} g_k \left[\frac{\partial}{\partial g_k} f(\mathbf{g}, \mathbf{h}) - \frac{\partial}{\partial h_k} f(\mathbf{g}, \mathbf{h}) \right] \\ & + \beta N \sum_{k=1}^{\infty} h_k \left[\frac{\partial}{\partial g_k} f(\mathbf{g}, \mathbf{h}) - \frac{\partial}{\partial h_k} f(\mathbf{g}, \mathbf{h}) \right]. \end{aligned} \quad (87)$$

The following theorem applies the operator semigroup to provide the mean-field limiting process $\{\mathbf{U}(t), t \geq 0\}$ for the sequence $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes, and indicates that this sequence of Markov processes asymptotically approaches a single trajectory identified by the unique and global solution to the system of mean-field equations. This proof is omitted here. Readers may refer to Li et al. [30] for more details.

Theorem 4 *Let f be continuous functions $f : \tilde{\Omega} \rightarrow \mathbf{C}^1$. Then for any $t > 0$*

$$\lim_{N \rightarrow \infty} \sup_{(\mathbf{g}, \mathbf{h}) \in \Omega_N} |\mathbf{T}_N(t)f(\mathbf{g}, \mathbf{h}) - f(\mathbf{u}(t, \mathbf{g}, \mathbf{h}))| = 0.$$

The convergence is uniform in $t \in [0, a]$ for any $a > 0$.

Finally, we provide some interpretation on Theorem 4. If $\lim_{N \rightarrow \infty} \mathbf{U}^{(N)}(0) = \mathbf{u}(0) = (\mathbf{g}, \mathbf{h}) \in \Omega$ in probability, then Theorem 4 shows that $\mathbf{U}(t) = \lim_{N \rightarrow \infty} \mathbf{U}^{(N)}(t)$ is concentrated on the trajectory $\Gamma_{\mathbf{g}} = \{\mathbf{u}(t, \mathbf{g}, \mathbf{h}) : t \geq 0\}$. This indicates the functional strong law

of large numbers for the time evolution of the fraction of each state of this supermarket model, thus the sequence $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes converges weakly to the expected fraction vector $\mathbf{u}(t, \mathbf{g}, \mathbf{h})$ as $N \rightarrow \infty$, that is, for any $T > 0$

$$\lim_{N \rightarrow \infty} \sup_{0 \leq s \leq T} \|\mathbf{U}^{(N)}(s) - \mathbf{u}(s, \mathbf{g}, \mathbf{h})\| = 0 \text{ in probability.}$$

Note that the limits are necessary for using the stationary probabilities of the limiting process to give an effective approximate performance of this supermarket model.

The Proof of Theorem 2

In the remainder of this Appendix, we discuss some useful limits of the fraction vector $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$ whose purpose is to give the proof of Theorem 2.

The following theorem gives the limit of the vector $\mathbf{u}(t, \mathbf{g}, \mathbf{h})$ as $t \rightarrow +\infty$, that is,

$$\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g}, \mathbf{h}) = \lim_{t \rightarrow +\infty} \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t, \mathbf{g}, \mathbf{h}).$$

This proof is omitted here. Readers may refer to Li et al. [30] for more details.

Theorem 5 *If $\tilde{\rho} < 1$, then for any $(\mathbf{g}, \mathbf{h}) \in \Omega$*

$$\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g}, \mathbf{h}) = \pi.$$

Furthermore, there exists a unique probability measure φ on Ω , which is invariant under the map $(\mathbf{g}, \mathbf{h}) \mapsto \mathbf{u}(t, \mathbf{g}, \mathbf{h})$, that is, for any continuous function $f : \Omega \rightarrow \mathbf{R}$ and $t > 0$

$$\int_{\Omega} f(\mathbf{g}, \mathbf{h}) d\varphi(\mathbf{g}, \mathbf{h}) = \int_{\Omega} f(\mathbf{u}(t, \mathbf{g}, \mathbf{h})) d\varphi(\mathbf{g}, \mathbf{h}).$$

Also, $\varphi = \delta_{\pi}$ is the probability measure concentrated at the fixed point π .

The following theorem indicates the weak convergence of the sequence $\{\varphi_N\}$ of stationary probability distributions for the sequence $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes to the probability measure concentrated at the fixed point π . This proof is omitted here. Readers may refer to Li et al. [30] for more details.

Theorem 6 (1) *If $\tilde{\rho} < 1$, then for a fixed number $N = 1, 2, 3, \dots$, the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ is positive recurrent, and has a unique invariant distribution φ_N .*

(2) *$\{\varphi_N\}$ weakly converges to δ_{π} , that is, for any continuous function $f : \Omega \rightarrow \mathbf{R}$*

$$\lim_{N \rightarrow \infty} E_{\varphi_N} [f(\mathbf{g}, \mathbf{h})] = f(\pi).$$

Based on Theorems 5 and 6, we obtain a useful relation as follows

$$\lim_{t \rightarrow +\infty} \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t, \mathbf{g}, \mathbf{h}) = \lim_{N \rightarrow \infty} \lim_{t \rightarrow +\infty} \mathbf{u}^{(N)}(t, \mathbf{g}, \mathbf{h}) = \pi.$$

Therefore, we have

$$\lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} \mathbf{u}^{(N)}(t, \mathbf{g}, \mathbf{h}) = \pi.$$

Clearly, the above analysis completes the proof of Theorem 2.

References

- [1] I.J.B.F. Adan, O.J. Boxma and J.A.C. Resing (2001). Queueing models with multiple waiting lines. *Queueing Systems* **37**, 65–98.
- [2] A. Aissani and J.R. Artalejo (1998). On the single server retrial queue subject to breakdowns. *Queueing Systems* **30**, 309–321.
- [3] S. Andradottir, H. Ayhan and D.G. Down (2007). Compensating for failures with flexible servers. *Operations Research* **55**, 753–768.
- [4] F. Baccelli, F.I. Karpelevich, M.Y. Kelbert, A.A. Puhalskii, A.N. Rybko and Y.M. Suhov (1992). A mean-field limit for a class of queueing networks. *Journal of Statistical Physics* **66**, 803–825.
- [5] M. Bramson, Y. Lu and B. Prabhakar (2010). Randomized load balancing with general service time distributions. In: *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 275–286.
- [6] M. Benaïm and J.Y. Le Boudec (2008). A class of mean-field interaction models for computer and communication systems. *Performance Evaluation* **65**, 823–838.
- [7] K.A. Borovkov (1998). Propagation of chaos for queueing networks. *Theory of Probability & Its Applications* **42 (No. 3)**, 385–394.
- [8] C. Bordenave, D. McDonald and A. Proutiere (2010). A particle system in interaction with a rapidly varying environment: mean-field limits and applications. *Networks and Heterogeneous Media* **5**, 31–62.

- [9] M. Bramson, Y. Lu and B. Prabhakar (2012). Asymptotic independence of queues under randomized load balancing. *Queueing Systems* **71**, 247–292.
- [10] M. Bramson, Y. Lu and B. Prabhakar (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *The Annals of Applied Probability* **23**, 1841–1878.
- [11] M.F. Chen (2004). *From Markov Chains to Non-Equilibrium Particle Systems*. World Scientific.
- [12] D.A. Dawson (1983). Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *Journal of Statistical Physics* **31**, 29–85.
- [13] N.G. Duffield (1992). Local mean-field Markov processes: An application to message-switching networks. *Probability Theory and Related Fields* **93**, 485–505.
- [14] K.R. Duffy (2010). Mean field Markov models of wireless local area networks. *Markov Processes and Related Fields* **16**, 295–328.
- [15] A. Economou and S. Kantaa (2008). Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters* **36**, 696–699.
- [16] S.N. Ethier and T.G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. John Wiley & Sons.
- [17] D. Fiems, T. Maertens and H. Brunee (2008). Queueing systems with different types of interruptions. *European Journal of Operational Research* **188**, 838–845.
- [18] N. Gast and B. Gaujal (2010). A mean eld model of work stealing in large-scale systems. *ACM SIGMETRICS Performance Evaluation Review* **38**, 13–24.
- [19] N. Gast and B. Gaujal (2011). A mean field approach for optimization in discrete time. *Discrete Event Dynamic Systems* **21**, 63–101.
- [20] N. Gast, B. Gaujal and J.Y. Le Boudec (2012). Mean-field for Markov decision processes: from discrete to continuous optimization. *IEEE Transactions on Automatic Control* **57**, 2266–2280.
- [21] C. Graham (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probabability* **37**, 198–201.

- [22] C. Graham (2004). Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probability Theory Related Fields* **131**, 97–120.
- [23] F. Kamoun (2008). Performance analysis of a non-preemptive priority queueing system subjected to a correlated Markovian interruption process. *Computers & Operations Research* **35**, 3969–3988.
- [24] C. Kipnis and C. Landim (2013). *Scaling Limits of Interacting Particle Systems*. Springer
- [25] A. Krishnamoorthy, P.K. Pramod and S.R. Chakravarthy (2012). Queues with interruptions: a survey. *Top* **22**, 290–320.
- [26] V.G. Kulkarni and B.D. Choi (1990). Retrial queues with server subject to breakdowns and repairs. *Queueing Systems* **7**, 191–208
- [27] Q.L. Li (2010). *Constructive Computation in Stochastic Models with Applications: The RG-Factorizations*. Springer and Tsinghua Press.
- [28] Q.L. Li (2014). Tail probabilities in queueing processes. *Asia-Pacific Journal of Operational Research* **31** (No. 2), 1–31.
- [29] Q.L. Li (2016). Nonlinear Markov processes in big networks. *Special Matrices* **4** (No. 1), 202–217.
- [30] Q.L. Li, G. Dai, J.C.S. Lui and Y. Wang (2014). The mean-field computation in a supermarket model with server multiple vacations. *Discrete Event Dynamic Systems* **24**, 473–522.
- [31] Q.L. Li, Y. Du, G. Dai and M. Wang, (2015). On a doubly dynamically controlled supermarket model with impatient customers. *Computers & Operations Research* **55**, 76–87.
- [32] Q.L. Li and J.C.S. Lui (2016). Block-structured supermarket models. *Discrete Event Dynamic Systems* **26**, 147–182.
- [33] Q.L. Li and F.F. Yang (2015). Mean-field analysis for heterogeneous work stealing models. In: *Information Technologies and Mathematical Modelling: Queueing Theory and Applications*, Springer, pp. 28–40.

- [34] Li Q.L., Y. Ying and Y.Q. Zhao (2006). A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research* **141**, 233–270.
- [35] T. Liggett (2012). *Interacting Particle Systems*. Springer
- [36] W. Li, D.H. Shi and X. Chao (1997). Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations. *Journal of Applied Probability* **34**, 546–555.
- [37] M. Luczak and C. McDiarmid (2006). On the maximum queue length in the supermarket model. *The Annals of Probability* **34**, 493–527.
- [38] M. Luczak and C. McDiarmid (2007). Asymptotic distributions and chaos for the supermarket model. *Electronic Journal of Probability* **12**, 75–99.
- [39] J.B. Martin (2001). Point processes in fast Jackson networks. *The Annals of Applied Probability* **11**, 650–663.
- [40] J.B. Martin and Y.M. Suhov (1999). Fast Jackson networks. *The Annals of Applied Probability* **9**, 854–870.
- [41] S.E. Martonosi (2011). Dynamic server allocation at parallel queues. *IIE Transactions* **43**, 863–877
- [42] I.L. Mitrany and B. Avi-Ttzhak (1968). A many server queue with service interruptions. *Operations Research* **16**, 628–638.
- [43] M.D. Mitzenmacher (1996). The power of two choices in randomized load balancing. PhD thesis, Department of Computer Science, University of California at Berkeley, USA.
- [44] M.D. Mitzenmacher, A. Richa and R. Sitaraman (2001) The power of two random choices: A survey of techniques and results. In: *Handbook of randomized computing* **1**, pp. 255–312.
- [45] A. Mukhopadhyay, A. Karthik, R.R. Mazumdar and F. Guillemin (2015). Mean field and propagation of chaos in multi-class heterogeneous loss models. *Performance Evaluation* **91**, 117–131.
- [46] M.F. Neuts and D.M. Lucantoni (1979). A Markovian queue with N servers subject to breakdowns and repairs. *Managment Scicens* **25**, 849–861.

- [47] R. Núñez-Queija (2000). Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems* **34**, 351–386.
- [48] R. Ravid, O. J. Boxma and D. Perry (2013). Repair systems with exchangeable items and the longest queue mechanism. *Queueing Systems* **73**, 295–316.
- [49] S. Saghafian, M.P. Van Oyen and B. Kolfal (2011). The “W” network and the dynamic control of unreliable flexible servers. *IIE Transactions* **43**, 893–907.
- [50] F. Spitzer (1970). Interaction of Markov processes. *Advances in Mathematics* **5**, 246–290.
- [51] Y.M. Suhov and N.D. Vvedenskaya (2002). Fast Jackson networks with dynamic routing. *Problems of Information Transmission* **38**, 136–153.
- [52] A. Sznitman (1989). *Topics in Propagation of Chaos*. Springer-Verlag, pp. 165–251.
- [53] S.R.E. Turner (1996). Resource pooling in stochastic networks. Ph.D. Thesis, Statistical Laboratory, Christ’s College, University of Cambridge.
- [54] S.R.E. Turner (1998). The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences* **12**, 109–124.
- [55] N.D. Vvedenskaya, R.L. Dobrushin and F.I. Karpelevich (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmissions* **32**, 15–27.
- [56] N.D. Vvedenskaya and Y.M. Suhov (1997). Dobrushin’s mean-field approximation for a queue with dynamic routing. *Markov Processes and Related Fields* **3**, 493–526.